Psychometrika

CONTENTS

(cont.)

VOLUME TWENTY-THREE DECEMBER 1958 NUMBER 4

Psychometrika

CONTENTS (Cont.)

VOLUME TWENTY-THREE DECEMBER 1958 NUMBER 4

PSYCHOMETRIKA—VOL. 23, NO. 4
DECEMBER, 1958

THE MYSTERY OF THE MISSING CORPUS*

FREDERICK MOSTELLER
HARVARD UNIVERSITY

Until recently, two happy vices stole a good share of my time. Of those burglars, the first was the reading of science-fiction stories, a source of entertainment that I discovered at the age of eight, but kept manfully in check because I never stumbled upon a continuous supply. Of late, however, real life has mocked this branch of fiction so closely that I now miss in it the elements of fantasy and escape that used to delight. You might say that, being crowded out by the real thing, we armchair spacemen face technological obsolescence.

My second time-stealer has been "Murder for Pleasure," as Howard Haycraft calls the detective story. Fortunately, I did not discover its existence until my twenties, so I had a little time for study before then. The great value of the detective story, quite apart from its educational aspects—after all, where else can one gather so much data about useful everyday questions such as tasteless poisons, homicidal law, and criminal psychology—is the fine training one gets in applied psychology and in the logic of everyday life.

Sigmund Freud and Carl Gustav Jung can look to their laurels because, after even a short course, we mystery-story fans can twist the slightest tongue-slip into a full confession, and impractical logic-choppers like Alfred North Whitehead and Bertrand Russell are pitifully pedestrian in their thinking, compared to the agile mental leaps we make—alongside Sherlock Holmes. We have to jump because Sherlock has a great reluctance to share his evidence with the reader. In fact, to put it bluntly, Sherlock cheated, but we excuse him because he performed before the rules were written. Now that we have rules of evidence for modern detective stories, all readers find it easy to search out the killer. (Anyway all readers I've ever met.) Indeed Robert Louis Stevenson commented upon this remarkable ability, "It is the difficulty of the police romance that the reader is always a person of such vastly greater ingenuity than the writer." And I personally have a sneaking suspicion that this is also one of the difficulties of Presidential Addresses. In best American style I wish to capitalize on my own deficiencies and take advantage of your ingenuity by pointing to a number of puzzles that seem to me to need more detective work in this Society—some minor like parking

*Presidential address delivered to the Psychometric Society, Washington, D. C., September 2, 1958.

violations and normality, others major like murder and scales of measurement.

Let us begin with a minor problem entitled in true mystery-story language by our great detective Thurstone: the Case of the Law of Comparative Judgment. When I talk with others about Thurstone's method of paired comparisons or one of the related scaling methods, several questions arise: What about the need for normality and for an underlying psychological continuum? Students ask why I think that such a complicated theory can possibly apply to attitudes, opinions, and psychophysical judgments.

To take up the last matter first, there is more than one attitude that can be taken toward these scaling methods. One attitude is that implied by the question—that the particular psychometric method used is a strong psychological theory taken quite seriously as such by its users. But an alternative attitude is that we have a statistical technique that makes it possible to summarize complex data. After all, what we like about paired comparisons is that, compared to other techniques, the judgments required seem rather simple and basic. What we don't like is the substantial body of data available at the end. Without further interpretation, if there are say 7 stimuli, we have $(7 \times 6)/2 = 21$ percentages to explain or to remember or $21n$ judgments, where n is the number of subjects. This is a great deal of information to keep organized. We want some way to summarize it. If we can replace the 21 numbers by one number for each stimulus, and if, further, we provide a rule by which the original percentages can be accurately recaptured from the shorter series of numbers, our summary has not lost any information. In this sense then it is possible to view the method of paired comparisons as a device for creating descriptive statistics, a method like the fitting of a regression plane that requires relatively little psychological theory for its justification. One is not forced to adopt this empirical attitude toward psychometric methods, but it sometimes helps, and its existence should be known.

Let us turn to the Case of the Eerie Curves on the Phantom Continuum, or Does Normality Really Matter?

In the original development of the law of comparative judgment the normal law was chosen, I presume, almost as a routine matter. Though the theory is developed around the distribution of responses to single stimuli, such responses are not actually used in the analysis. By distribution of responses I mean the distribution of the phantom random variable whose value on the conjectured psychological continuum is assumed covertly in the organism when one stimulus is presented. You can easily see why there is a relationship to the detective story here.

The family of distributions that is used in the analysis is not the family corresponding to judgments of original "measurements" or "response values" but that corresponding to the judgments of the difference of two such response values. If we actually had the distribution of the difference of two

independent random variables, identically distributed, except possibly for slippage, then we could guarantee a symmetrical distribution; and insofar as the original measurements might have been approximately normal, the all-important distribution of differences would be more nearly normal (in the sense of cumulants). Such reasoning might be used to justify the normality assumption as a close approximation, but it is tenuous reasoning at best.

Why is it tenuous? First, the independence assumption is not very appropriate, for reasons I will not give in detail. But, briefly, since the two stimuli are presented at about the same time, we expect the subject to be much more like himself in that little interval than like someone else at another time, or even than like himself at a distant time, and thus we do expect correlations. Second, we are not even dealing with an actual difference of random variables, but with single judgments about a difference. And there is nothing in the theory that pushes this distribution toward normality. We cannot therefore depend upon the previous train of reasoning to establish solidly the shape of the response curves or even to say much about the family of curves we use. On the other hand, maybe we can say that some family works well.

Some kind of curve or family of curves is needed to grade the response percentages. Just how important the shape of this curve is has never been made clear. To put it another way, except for some work of Hull and his associates, there does not seem to be much research either empirical or theoretical studying the impact of various shapes upon goodness of fit. The behavior of the tails of the distribution could be important—though we don't usually give much weight to the very extreme proportions like .999 or .001. Emphasis on the shape of tails is due to a rather practical remark of the late Charles Winsor; he pointed out that most distributions met in practice are approximately "normal in the middle."

Perhaps a word more about the nature of the curves under discussion would not be amiss. We usually talk about the normal as if we were concerned with a frequency distribution. This notion arises from the model used to derive the curve. An alternative view is that we are dealing with the equivalent of a dosage response curve or an operating characteristic, and that as we move from left to right (increase the dose), the curve shows the percent that die from the administration of the poison. To push the matter further, some poisons have response curves that increase as we move from left to right, reach a plateau, and then decrease—a thing a cumulative distribution does not do. I bring this out to emphasize that the interpretation of a distribution is not absolutely essential rather than to suggest the use of pathological functions for grading responses when we are doing well without them.

Without trying to carry out the research needed to discover the sensi-

TABLE 1A

Scale Values for Vegetable Example Using a Variety of Scaling Devices

	Turnips	Cabbage	Beets	Asparagus	Carrots	Spinach	String beans	Peas	Corn
Uniform	-.321	-.167	-.118	-.007	.042	.050	.140	.158	.222
Arc sin $\sqrt{p}$	-20.80	-10.20	-7.28	-.27	2.73	3.26	8.85	9.87	13.84
Case V	-.988	-.465	-.333	-.008	.129	.156	.412	.456	.642
$\frac{1}{2} e^{-\lvert x \rvert}$	-1.194	-.502	-.365	.010	.156	.194	.472	.509	.719
t_{10}	-1.051	-.487	-.349	-.007	.137	.167	.435	.480	.676

TABLE 1B

Scale Values for Table 1A Adjusted to Give Turnips the Position 0 and Corn the Position 1

	Turnips	Cabbage	Beets	Asparagus	Carrots	Spinach	String beans	Peas	Corn
Uniform	0	.284	.374	.578	.669	.683	.849	.882	1
Arc sin $\sqrt{p}$	0	.306	.390	.593	.679	.695	.856	.885	1
Case V	0	.321	.402	.601	.685	.702	.859	.886	1
$\frac{1}{2} e^{-\lvert x \rvert}$	0	.362	.433	.629	.706	.726	.871	.890	1
t_{10}	0	.327	.406	.605	.688	.705	.860	.887	1

TABLE 2

Correlations between Scale Positions by Various Methods

	Uniform	Arc sin $\sqrt{p}$	Case V	$\frac{1}{2} e^{-\lvert x \rvert}$	t_{10}
Uniform	1.				
Arc sin $\sqrt{p}$	.9997	1.			
Case V	.9993	.9999	1.		
$\frac{1}{2} e^{-\lvert x \rvert}$	.9965	.9982	.9990	1.	
t_{10}	.9990	.9998	.99998	.9992	1.

TABLE 3

Comparison of Recaptured Proportions and Observed for the "Carrot" Observations

	Turnips	Cabbage	Beets	Asparagus	Carrots	Spinach	String beans	Peas	Corn	**
Observed	.122	.257	.264	.439	(.5)	.493	.574	.709	.764	
Uniform	.137	.291	.340	.451	(.5)	.508	.598	.616	.680	.353
Arc sin $\sqrt{p}$	.134	.282	.329	.448	(.5)	.509	.606	.623	.689	.320
Case V	.132	.276	.322	.446	(.5)	.511	.611	.628	.696	.298
$\frac{1}{2} e^{-\lvert x \rvert}$	.130	.259	.298	.432	(.5)	.519	.636	.649	.716	.247
t_{10}	.131	.293	.319	.444	(.5)	.512	.614	.630	.699	.308

** Sum of absolute deviations from observed values.

tivity of the method to the curve, I shall give a notion of what can happen. Table 1A shows the fitted values provided by a variety of different curves used to grade the response percentages in an example of vegetable preferences shown by Guilford [2]. The curves used were the cumulative distributions: for the uniform on the interval 0 to 1, for the arc sin $\sqrt{p}$ on the interval 0° to 90°, for the normal, $\frac{1}{2}e^{-|x|}$, and for the t-distribution with 10 d.f. (the 10 was chosen without any particular idea in mind). The order of listing of the distributions is that of the height of the tails. Clearly the uniform distribution has the lowest tails, arc sin $\sqrt{p}$ a little higher, and so on. (The densities should be thought of as distributed over the whole real line, and so the tails of the uniform have zero height.) In Table 1B we have transformed linearly so that the scale values run from 0 for the lowest to 1 for the highest stimulus. These transformed values show that the scale values are very similar.

Table 2 shows this similarity by giving the correlations between stimulus positions assigned by the various curves. Unless you are used to looking at correlations between one set of numbers and a mild transformation of the set, you will regard these correlations as rather high.

While the spacings are roughly invariant except for linear transformations, the reproduction of the original percentages is not the same; these reproductions are shown in Table 3 for one line of the original table. Note that the sum of the absolute errors decreases as we move down the rows and then increases. Appropriate comparisons should, of course, be made for the whole table. I might mention that the behavior exhibited here agrees with my other experiences. The uniform and the arc sine were not as good as the normal, and the normal is not quite as good as something with higher tails.

A useful piece of research, then, would be to explore the sensitivity of the method of paired comparisons to the shape of the curve used to grade the responses.

Next we have the Case of the Unconfused Judges. We need, too, a rethinking of the theory of the paired comparison approach and its relatives, expecially in examples where the stimuli are recognizable by individuals and are ordered rather consistently by them. Once, following an approach suggested by M. G. Kendall [3], I explored the intransitivities within individuals in a set of paired comparison items in an opinion poll. I soon became weary of the task because so few people made a mistake. Indeed, the Kendall development is intended to test the nullest of null hypotheses—that everyone is guessing randomly, and therefore that there is no order. While there may be circumstances under which this is a sensible null hypothesis or at least a good baseline to measure from, in situations where stimuli are easily recognized, distinguished, and familiar, it is not likely to be appropriate.

If there is practically no confusion within respondents in the method of paired comparisons, this suggests strongly that the stimuli are widely spaced

as far as individuals are concerned. If so, we could, by examining the responses of individuals in detail, capture the ranking proposed by each individual. Then we could use as the scale value for a stimulus the average rank assigned. This sounds like a lot of work, but when there is *no* inconsistency within an individual it is possible to get the average ranks directly from the paired comparison table quite easily. We just sum the fractions in a column, including a 1 rather than a 0.500 in the main diagonal. This extra 1 adjusts so that we rank from 1 to N (the number of stimuli) rather than 0 to N–1. Now if we have slight inconsistencies this method will give an approximate ranking.

TABLE 4

Formation of Paired Comparisons Table from the Rankings *

Order	Ranking	Number choosing order
1	ABC	n_1
2	ACB	n_2
3	BAC	n_3
4	BCA	n_4
5	CAB	n_5
6	CBA	n_6
Rank	321	

	Stimulus		
	A	B	C
A	Σn_i	$n_3+n_4+n_6$	$n_4+n_5+n_6$
B	$n_1+n_2+n_5$	Σn_i	$n_2+n_5+n_6$
C	$n_1+n_2+n_3$	$n_1+n_3+n_4$	Σn_i
Sum	$3n_1+3n_2+2n_3$ $+n_4+2n_5+n_6$	$2n_1+n_2+3n_3$ $+3n_4+n_5+2n_6$	$n_1+2n_2+n_3$ $+2n_4+3n_5+3n_6$

* Entries in paired comparison table should be divided by $n = \Sigma n_i$ to obtain proportions rather than frequencies.

Table 4 shows the method for three stimuli. There n_i is the number of individuals responding in a particular manner; A, B, and C are labels for the stimuli; the ranks are 3, 2, and 1. The paired comparison table shows in the cell the total number of people preferring the stimulus at the head of its column to that listed at the left of its row. It is readily verified that the sum of the ranks for the stimuli is identical to the sums at the foot of the paired comparison table.

There is an essential identity between this technique and the use of the cumulative uniform distribution as the dosage response curve for paired comparisons. In this interpretation, however, what is basic is the number of individuals in the sample or population who choose each of the $N!$ arrangements, whereas in the usual analysis of paired comparisons it is the $N(N-1)/2$ preference percentages. In the paired comparisons approach we obtain a set of $N(N-1)/2$ percentages, replace these by a set of N numbers. Then we use

differences of these N numbers (in Case V) together with a normal table or other rule to approximately recapture the percentages. In the technique using the average ranking, there is no way to recapture from the percentages the $N!$ basic numbers (those choosing the several orders), nor do I see the proper analogue to the recapture of the percentages.

Is there a natural sequence of paired comparison models that depends upon the extent of confusion within individuals? Or is that what the usual technique is? The question needs investigation and explication. Gerard and Shapiro [1] have suggested a method for assessing the degree of confusion.

Before leaving this topic I should mention that this is by no means the only technique for handling these rankings or even the confused rankings. Indeed, a few years age a president of the Society, a namesake of the great scientific detective created by R. Austin Freeman, proposed a technique for factoring, or should I say fracturing, a set of individuals into sensible subgroups on the basis of multiple measurements.

These problems are relatively minor compared to The Tragedy of y, or The Mystery of the Missing Corpus. By y, I mean the psychometric measures, and the missing corpus is the main body of the literature. We are now in the 23rd volume of *Psychometrika*. Yet we still do not have a general theory or set of theories of scaling—a set of theories that relates the various psychometric methods to one another, and that relates psychological continua to physical continua. I regard the development of such theories as a very crying need. I do not anticipate that the matter will be as simple as the creation of a single all-embracing theory, nor do I think that a mathematical approach alone will be of value. We need an extensive experimental program as well, covering a variety of physical continua, of attitude and opinion material, and employing a variety of measurement techniques.

We have recently seen the interest of S. S. Stevens in the relationship between psychological and psychophysical continua. He finds by the use of magnitude estimation and production that the relationship between certain physical and psychological continua is not logarithmetic in character as Fechner's law suggests but a power law ($y = ax^b$, where y is the value on the psychological continuum and x the value on the physical continuum). He has used his results to argue that certain methods of measurement are less appropriate than others.

If we had good theory to parallel good experimentation, we could more easily agree on the appropriateness or inappropriateness of certain methods of scaling, either in relation to certain physical continua or in relation to one another. For example, we want to know the relations between various psychometric methods. A good theory ought either to tell us how to relate the various psychometric methods or, by bringing out the incompatible elements of the different methods, tell us how inappropriate it is to attempt to find a relationship.

In addition to the intensive experimental work recently done by Stevens and his coworkers, there has been some instructive work in a mathematical direction by R. Duncan Luce. This work considers the two more important measurement scales as classified by Stevens.(As mystery-story fans you will want to be assured that this is the same Stevens we talked about before and not his long-lost Australian twin brother. It is.) These two scales are the interval and the ratio scale. The first is essentially determined up to linear transformations, that is, you are free to choose an origin and a unit of measurement arbitrarily, while the second or ratio scale leaves you free only to choose the unit of measurement.

Luce inquires what laws could relate an interval or ratio scale corresponding to a physical continuum with an interval or ratio scale corresponding to a psychological continuum. (There are four possible pairings: interval-interval, interval-ratio, ratio-interval, ratio-ratio.) As a lever for this consideration he makes the innocuous assumption that admissible transformations on one continuum ought to produce only admissible transformations on the other and that insofar as the choice of origin or of unit of measurement is arbitrary on the two scales, it should continue to be arbitrary. That is, it should not be forced as a result of the transformation. I shall not give all of Luce's results, but I mention two. In order to have a ratio scale on both, the power law is the only possible psychophysical relationship. A little more usual situation with a ratio scale on the physical continuum and an interval scale on the psychological continuum leads either to a linear transformation of the logarithm of the physical continuum or to a linear transformation of an arbitrary power of the physical continuum (a slight generalization of the power law in that it allows the addition of a constant).

These results are striking moves in one of the directions where research is needed. On the other hand, as a solution to the mystery story, they give one the cold shivers. Somehow the theories of physics are not so restrained in their formulas. How can this be? There are several matters that need clarification. First, there is often more than one variable involved in the relationship. The existence of additional variables would broaden the admissible functions. A few possibilities are: dependence upon the number of categories used, upon some aspects of a physical stimulus, or upon features of the instructions.

A second place to look for expansion of possible laws is in the matter of the admissible transformations themselves. The two under consideration, the general linear transformation and the more restrictive multiplicative transformation, do not begin to exhaust the possible transformations even in the one-dimensional case. We are familiar, besides the nominal and ordinal classes, with simple translations and with the more complicated type of scale generated by the Coombs procedure, which he calls an ordered metric. But many classes of transformations could be used that most of us have

never considered. And when I think, in addition, of the possibility of a slight rumpling of continua owing to chance variation in functions, I begin to be a bit afraid of the strict interpretation of interval and ratio scales in the psychological experiments with which I am familiar. Nevertheless, I feel that these scales are instructive and appropriate for the first investigations.

While we are on this topic, there is the related problem of just what kind of statistic is appropriate in the presence of various types of scales of measurement. And that question sets me to wondering just what kind of a scale of measurement is appropriate for test scores. One might think either of items that are scored one or zero or of items that are scored in some other manner. Does doubling the score mean doubling the knowledge?

If one knew one had exactly an ordinal scale, no more no less, then the notion that order statistics or percentiles are appropriate measures and that averages are not is well taken. But when I do not know exactly what kind of scale I have, for instance, I may have nearly an interval scale in the test score case, it seems sensible to use the statistics appropriate to the type of scale I think I am near. In taking such action we may find the justification vague and fuzzy. One reason for this vagueness is that we have not yet studied enough about classes of scales, classes appropriate to real life measurement, with perhaps real life bias and error variance. There is some danger in pushing for too tight a model of measurement. We need a little room for the actual variation of the laboratory that every experimenter recognizes. At the same time we hope the laboratory situation will measure ever more sharply.

Let us illustrate a little further the importance of a slightly soft theory. It will be recalled that we obtained very high correlations between the scale values assigned to the several vegetable stimuli by the different scaling techniques. These different techniques are essentially modest transformations of the scale. One might suppose that it is an accident that such high correlations are achieved but perhaps a further mathematical example will clarify the point. Suppose we consider the transformations

$$y = bx + (1 - b)x^2$$

for x running from 0 to 1 and for b taking a value between 0 and 1. With these restrictions this is a monotonic quadratic transformation that sends $x = 0$ into $y = 0$ and $x = 1$ into $y = 1$. If we assume the x's are uniformly distributed between 0 and 1, it is easy to compute the correlation between x and y. And after a little arithmetic, we find that the correlation for a given value of b is

$$\rho_b = \frac{1}{\sqrt{1 + \frac{(1 - b)^2}{15}}} \cong \frac{1}{1 + \frac{(1 - b)^2}{30}} \cong 1 - \frac{(1 - b)^2}{30}.$$

Thus in the worst case with $b = 0$, we get a correlation of approximately 0.97, or if you prefer to deal in squares of correlation coefficients, the value is exactly 15/16, or about 0.94. This shows that nonlinearly related scales can yield high correlations. And this supports the idea that we should not too lightly abandon a statistical method because we cannot assure ourselves that we do have exactly an interval or ratio scale.

An additional reason for vagueness is that we haven't tried to use the information contained in the purpose of making the measurements. Very often the whyness has a good deal to say about the choice of statistics, as you all know. But "why" is a matter that has not been too often raised in a discussion like the present one about measurement. The truth is that mathematics alone will take one a certain distance, just as logic will help the detective, but after a while the detective needs clues, and information about motives and opportunities, just as we need facts about and motives for the measurement.

I have tried to point to the need for broader theories relating methods of measurement, and I have pointed to the parallel need for experimental work. I shall look forward to your contributions to the search for the missing corpus in the journals.

REFFERENCES

[1] Gerard, H. B. and Shapiro, H. N. Determining the degree of inconsistency in a set of paired comparisons. *Psychometrika*, 1958, **23**, 33–46.
[2] Guilford, J. P. *Psychometric methods.* (2nd ed.) New York: McGraw-Hill, 1954.
[3] Kendall, M. G. *Advanced theory of statistics.* London: Lippincott, 1943.

Manuscript received 9/2/58

PSYCHOMETRIKA—VOL. 23, NO. 4
DECEMBER, 1958

SOME RELATIONS BETWEEN GUTTMAN'S PRINCIPAL COMPONENTS OF SCALE ANALYSIS AND OTHER PSYCHOMETRIC THEORY

FREDERIC M. LORD
EDUCATIONAL TESTING SERVICE

Guttman's principal components for the weighting system are the item scoring weights that maximize the generalized Kuder-Richardson reliability coefficient. The principal component for any item is effectively the same as the factor loading of the item divided by the item standard deviation, the factor loadings being obtained from an ordinary factor analysis of the item intercorrelation matrix.

By Guttman's definition, a principal component of the score system for a set of item responses consists of a vector of numerical scores, one for each examinee. There is a corresponding principal component of the weighting system consisting of a vector of numerical weights, one for each possible response to each test item. The system has the property that the numerical score for each examinee is proportional to the sum of the weights for the item responses selected by him ([9], pp. 315–321).

The first principal component constitutes the solution to the problem of assigning numerical scores to the examinees in such a fashion as to maximize a certain correlation ratio ([5], pp. 327 ff.). There are also second, third, and subsequent principal components, each of which corresponds to a *local* maximum for the correlation ratio, and each of which therefore satisfies the mathematical equations for a maximum. Each principal component of the weighting system is a latent vector of a matrix whose elements are certain "chi-square product-moments" ([5], p. 332). This is an $n \times n$ matrix, where n is the total number of different possible responses (e.g., if the test contains sixty 5-choice items, then $n = 60 \times 5 = 300$). The principal components of the scoring system may be obtained as the latent vectors of an $N \times N$ matrix, where N is the number of examinees, or they may be derived from the weights by simply adding together the weights of the responses selected by each examinee.

The present article will show the following.

1. *Guttman's principal components for the weighting system are the same as the sets of weights that will maximize the generalized Kuder-Richardson reliability coefficient.*

2. *Guttman's principal components for the weighting system* (and thus the scoring weights for maximizing test reliability, also) *are effectively the same as certain sets of item weights obtained by factoring the matrix of item*

intercorrelations. The weight for any item is equal to its unrotated factor loading divided by its standard deviation. The matrix factored is the $m \times m$ matrix of interitem product moment correlations, where m is the number of items; the Hotelling (principal components) method of factor analysis ([10], ch. 20) is used.

3. *Guttman's principal components for the scoring system are perfectly correlated with the scores obtained for the examinees when the items are weighted so as to maximize the generalized Kuder-Richardson reliability coefficient.*

These results are believed to be new. Guttman discusses Spearman-Thurstone factor analysis of the item intercorrelation matrix ([9], pp. 191–205) but does not mention the results presented here. His method of obtaining the principal components requires the factoring of a matrix with n/m times as many rows and columns as does the present method.

It will be convenient first to give a proof that the Kuder-Richardson reliability coefficient is maximized by a scoring formula that weights the standardized item scores by their factor loadings on the first factor. The relationship of Guttman's principal components to test reliability is treated in a second section.

Weighting Items to Maximize the Generalized Kuder-Richardson Reliability Coefficient

A generalization of the Kuder-Richardson formula–20 reliability coefficient, appropriate whenever the test items are to be scored with weights other than zero and one, has been described by Dressel [2] and by Cronbach [1], who suggests the more convenient term *coefficient alpha.* This coefficient is the same as a lower bound to the reliability, called L_3 , developed independently by Guttman [6]. The formula recommended by Hoyt and Stunkard [8] also can be shown to be mathematically identical with the others. In a recent article, Tryon [11] advances further lines of reasoning to justify this same coefficient.

The present section derives the scoring weights that will maximize coefficient alpha. No reliability coefficient that is actually computable will ever meet exactly the requirements that would be theoretically desirable in an ideal coefficient. Some objections have been raised against coefficient alpha. In particular, objection has been raised against any derivation that assumes, in effect, that all items are equally good measures of the same common factor. Various other derivations of this coefficient are given in the references cited [e.g., 11] however, and it seems likely that it will continue to play an important role in psychometric theory. Further discussion of this coefficient would be inappropriate here.

The general formula for coefficient alpha is

$$\alpha = \frac{m}{m-1}\left(1 - \frac{\sum_{i=1}^{m} V_i}{V_t}\right), \tag{1}$$

where m is the number of test items, V_t is the variance of the test scores, and V_i is the variance of the weighted scores on item i. Denote by s_{ij} the covariance between item i and item j *before* weighting the item scores, and denote by s_i^2 the variance of item i *before* weighting. Let W_i be the weight assigned to the score on item i, the test score of an examinee being the weighted sum of the scores on responses chosen by him. It is readily shown that, *after* weighting, $V_i = s_i^2 W_i^2$ and $V_t = \sum_i \sum_j s_{ij} W_i W_j$. Coefficient alpha may thus be written

$$\alpha = \frac{m}{m-1}\left(1 - \frac{\sum_i s_i^2 W_i^2}{\sum_i \sum_j s_{ij} W_i W_j}\right). \tag{2}$$

The present problem is as follows. Given the values of s_i^2 and of s_{ij} for a given test, find the scoring weights, W_i , that will maximize coefficient alpha. The value of α in (2) will remain unchanged if all the values of W_i ($i = 1, \cdots, m$) are multiplied by the same constant factor; hence, one restriction may be arbitrarily imposed on the values of W_i . It is convenient to require that the W_i shall be so restricted that the variance of test scores remains fixed:

$$V_t = \sum_i \sum_j s_{ij} W_i W_j = \text{constant}. \tag{3}$$

Since α can never be more than 1, the problem is seen to be one of maximizing one quadratic form, $\sum s_i^2 W_i^2$, when another, $\sum\sum s_{ij} W_i W_j$, is held constant. Denote the two quadratic forms in matrix notation by $w'D^2w$ and $w'Sw$, respectively, where w is the column vector $\{W_1, W_2, \cdots, W_m\}'$, D is the diagonal matrix whose elements are s_i , and $S = [s_{ij}]$. A general theorem on quadratic forms ([12], pp. 170–171) states that the values of W_i giving a maximum are those satisfying the matrix equation

$$(D^2 - \lambda S)w = 0, \tag{4}$$

where λ is a constant to be determined.

A more convenient equation is obtained by premultiplying (4) by $-\lambda^{-1}D^{-1}$ and making the substitutions $w = D^{-1}u$, $\mu = 1/\lambda$, and $D^{-1}SD^{-1} = R$:

$$(R - \mu I)u = 0. \tag{5}$$

Equation (5) is the characteristic equation of the matrix of interitem product moment correlations, $R = [r_{ij}]$. The desired optimum weights for maximizing α are thus

$$\hat{W}_i = U_i/s_i , \tag{6}$$

where $\{U_1, U_2, \cdots, U_m\}'$ is the first characteristic vector of the correlation matrix.

The foregoing result may be summarized in slightly different words by saying that *the test score having the maximum generalized Kuder-Richardson formula-20 reliability coefficient is obtained by a scoring formula that weights the standardized item scores by their factor loadings on their first principal component.* Here it is understood that attention is restricted to test scores that are linear functions of the item scores, that the "standardized item score" is the unweighted item score divided by s_i , and that the factor loadings are obtained by a (principal components) factor analysis of the interitem product moment correlations with unities in the diagonal.

The results given here are essentially the same as those obtained by Horst [7] for a different problem. He showed that the variance of composite score on a test battery is maximized (given a fixed value for the sum of squares of the weights) when the standardized score on each test is entered into the composite with a weight proportional to its factor loading on the first principal component of the battery. Edgerton and Kolbe [3] and Wilks [13] also reached essentially the same solution after starting with still different problems.

There are several other methods for determining optimum weights for combining tests into a composite ([4], ch. 20; [14]) but these require a knowledge of the reliability of each test. They are not relevant for the present problem where optimum weights are to be found for each test item without knowledge of any coefficient of item reliability.

*Relation of Guttman's Principal Components to the Generalized Kuder-Richardson Reliability Coefficient**

Let x_{ic} be the scoring weight of alternative response c for item i. Let y_{ia} be the score obtained by examinee a on item i, so that $y_{ia} = x_{ic}$ whenever examinee a chooses response c. Following Hoyt and Stunkard [8], let $y_{\cdot a} = \sum_{i=1}^{m} y_{ia}$ (the total score of examinee a), $y_{i\cdot} = \sum_{a=1}^{N} y_{ia}$, and $y_{\cdot\cdot} = \sum_i \sum_a y_{ia}$. An analysis of variance table may be written in part as shown below.

	Sum of squares
Among examinees	$A = \frac{1}{m} \sum_a y_{\cdot a}^2 - \frac{1}{Nm} y_{\cdot\cdot}^2$
Among items	$B = \frac{1}{N} \sum_i y_{i\cdot}^2 - \frac{1}{Nm} y_{\cdot\cdot}^2$
Residual	$C = T - A - B$
Total	$T = \sum_i \sum_a y_{ia}^2 - \frac{1}{Nm} y_{\cdot\cdot}^2$

*The writer's original mathematical proof restricted itself to the case of dichotomous items. The simpler proof presented here, valid for polychotomous items, was worked out subsequently by Professor Ledyard R Tucker.

Now, by definition,

$$V_i = \frac{1}{N}\sum_a y_{ia}^2 - \frac{1}{N^2}y_{i\cdot}^2 \,; \tag{7}$$

thus,

$$\sum_i V_i = (T - B)/N. \tag{8}$$

Likewise,

$$V_t = \frac{1}{N}\sum_a y_{\cdot a}^2 - \frac{1}{N^2}y_{\cdot\cdot}^2 = mA/N. \tag{9}$$

Consequently, from (1),

$$\alpha = \frac{m}{m-1}\left(1 - \frac{T-B}{mA}\right) = 1 - \frac{C}{(m-1)A}. \tag{10}$$

Guttman's principal components of the weighting system are the item scoring weights that maximize the correlation ratio ([5], eq. 1)

$$\eta_x^2 = \frac{A}{T}. \tag{11}$$

Since $T = A + B + C$,

$$\frac{1}{\eta_x^2} = 1 + \frac{B}{A} + \frac{C}{A}. \tag{12}$$

It only remains to show that the item scoring weights, $\hat{W}_i$, that maximize (11) are the same as those that maximize (10).

Guttman has shown ([5], eq. 16) that his optimum item weights have the property that the average score for an item ($y_{i\cdot}/N$) is the same for all items. (Since the origin is arbitrary, Guttman makes this average equal to zero.) Thus $B = 0$ whenever Guttman's optimum scoring weights are used. This condition on the item weights imposes no restriction on the value of α: it is well known that test reliability depends on the spread of the scoring weights assigned to the items, not on their average value.

When $B = 0$, it is seen from (12) and (10) that

$$\eta_x^2 = \frac{1}{1 + \frac{C}{A}} = \frac{1}{1 + (m-1)(1-\alpha)}. \tag{13}$$

It is obvious from (13) that η_x will be maximized by maximizing α, and conversely.

REFERENCES

[1] Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, **16**, 297–334.

[2] Dressel, P. L. Some remarks on the Kuder-Richardson reliability coefficient. *Psychometrika*, 1940, **5**, 305–310.

[3] Edgerton, H. A. and Kolbe, L. E. The method of minimum variation for the combination of criteria. *Psychometrika*, 1936, 1, 183–187.
[4] Gulliksen, H. *Theory of mental tests*. New York: Wiley, 1950.
[5] Guttman, L. The quantification of a class of attributes: a theory and method of scale construction. In P. Horst (Ed.), *The prediction of personal adjustment*. Soc. Sci. Res. Council, Bull. 48, 1941. Pp. 321–345.
[6] Guttman, L. A basis for analyzing test-retest reliability. *Psychometrika*, 1945, 10, 225–282.
[7] Horst, P. Obtaining a composite measure from a number of different measures of the same attribute. *Psychometrika*, 1936, 1, 53–60.
[8] Hoyt, C. J. and Stunkard, C. L. Estimation of test reliability for unrestricted item scoring methods. *Educ. psychol. Measmt*, 1952,12, 756–758.
[9] Stouffer, S. A. (Ed.) Measurement and prediction. *Studies in social psychology in World War II, Vol. IV*. Princeton, N. J. : Princeton Univ. Press, 1950.
[10] Thurstone, L. L. *Multiple-factor analysis*. Chicago: Univ. Chicago Press, 1947.
[11] Tryon, R. C. Reliability and behavior domain validity: reformulation and historical critique. *Psychol. Bull.*, 1957, 54, 229–249.
[12] Turnbull, H. W. and Aitken, A. C. *An introduction to the theory of canonical matrices*. Toronto: Blackie, 1950.
[13] Wilks, S. S. Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, 1938, 3, 23–40.
[14] Woodbury, M. A. and Lord, F. M. The most reliable composite with a specified true score. *Brit. J. statist. Psychol.*, 1956, 9, 21–28.

Manuscript received 9/11/57
Revised manuscript received 2/11/58

PSYCHOMETRIKA—VOL. 23, NO. 3
SEPTEMBER, 1958

TO WHAT EXTENT CAN COMMUNALITIES REDUCE RANK?*

LOUIS GUTTMAN
THE ISRAEL INSTITUTE OF APPLIED SOCIAL RESEARCH AND
THE HEBREW UNIVERSITY IN JERUSALEM

The question is raised as to whether the null hypothesis concerning the number of common factors underlying a given set of correlations should be that this number is small. Psychological and algebraic evidence indicate that a more appropriate null hypothesis is that the number is relatively large, and that smallness should be but an alternative hypothesis. The question is also raised as to why approximation procedures should be aimed primarily at the observed correlation matrix R and not at, say, R^{-1}. What may be best for R may be worst for R^{-1}, and conversely, yet R^{-1} is directly involved in problems of multiple and partial regressions. It is shown that a widely accepted inequality for the possible rank to which R can be reduced, when modified by communalities, is indeed false.

When Charles Spearman hypothesized, some fifty years ago, that correlations among certain mental test scores could be accounted for by but a single common factor, this was greeted by many of his colleagues with substantial scepticism. In current terminology, initiated by L. L. Thurstone, they thought it implausible that communalities could be found that would reduce the given type of observed correlation matrix to rank one. Thurstone later hypothesized that relatively small rank could be attained for correlation matrices of mental test data by use of communalities. This hypothesis, too, has encountered a measure of disbelief in various quarters.

Motivation for seeking small rank stems from the desire to reproduce the observed correlations among n variables by using scores on a smaller number, say m, of common factors. A necessary and sufficient condition that there exist m common factors to do the reproducing trick is that there exist communalities that reduce the rank of the observed correlation matrix to m, but leave the matrix Gramian [8]. In this sense, *rank* and *number of common factors* are equivalent. It is algebraically convenient to deal with rank in studying the possible number of common factors for a given set of data.

Empirical attempts to estimate minimal rank in given cases have hitherto not been clear cut for lack of rigorous theory and computing routines for fallible data. Among the better efforts in this direction are the works of Lawley [13] and Rao [16]. But these do not presume to do more than give a lower bound to the minimal rank; they do not provide any upper bounds.

*This research was facilitated by a noncommitted grant-in-aid to the writer from the Ford Foundation.

Regardless, if all results published to date be taken at their face value, they together constitute abundant evidence against the Thurstone hypothesis. Hundreds of different common factors for mental abilities have been "identified" by Thurstonian computing routines, and the number is still growing, with no upper limit in sight (cf. [2] and the discussions in [12] and [19]).

The association of the notion of communality coefficients with that of minimal rank seems to be an historical accident, as it is not logically necessary [4, 8, 9, 10]. Identifying the concept of small rank with that of scientific parsimony also seems fortuitous, as this too has no logical compulsion; other kinds of parsimony are possible [6, 12]. In view of the evidence favoring large rank—even when communalities are used—for the entire domain of mental abilities, it is fortunate for the communality concept that it is useful and meaningful regardless. For example, a unique definition of communalities is sometimes possible in terms of image analysis, without considering rank at all [4, 9, 10]. There are other possibilities for other cases, especially where facet design is used for the tests [12]. Communalities can lead to parsimonious descriptions of data without reference to rank.

Individual research projects, however, typically study but a small sector of the domain of mental abilities and not the entire domain. Is it still not true that "in general" communalities will meaningfully reduce rank in the subdomains? Also, isn't it an algebraic fact that, psychological meaning aside, communalities for a given correlation matrix of order n can always be found to reduce the rank to m, where the inequality is satisfied,

$$m \leqq \tfrac{1}{2}(2n + 1 - \sqrt{8n + 1}) \tag{1}$$

as was shown by several authors [1, 14, 18]?

The purpose of the present paper is to explore these last questions. Despite the rather widespread belief in inequality (1), it can be proved false. Its authors make a valiant attempt to establish an algebraic upper bound to minimal rank, but introduce an important fallacy. In fact, the best possible universal upper bound to minimal m is $n - 1$, as shall be shown.

Further algebraic evidence will be submitted against small rank being the general algebraic case. For fallible data, this implies that large rank should be the null hypothesis, not to be rejected unless there is strong empirical evidence to the contrary. Most current computing routines have the opposite orientation; their justification accordingly may require reexamination. Because of the confusion as to the nature of the rank hypothesis to be tested by fallible data, we open our discussion of communalities with this problem of orientation. Psychological considerations will be stressed no less than purely algebraic ones.

Great Complexity as the Null Hypothesis for Mental Test Data

Empirical research of the past five decades and more has consistently revealed positive correlations of varying sizes among mental test scores. On the face of it, the structure of the interrelations seems very complex; psychologists will tend to demand substantial evidence before they will accept any hypothesis that states otherwise.

For example, no one will seriously entertain the simplest hypothesis of zero rank. Suppose some experimenter were to take some standard mental tests, administer them to 25 pupils, apply a standard test of significance to the sample correlations, and conclude that no population correlation coefficient was different from zero. The reaction of his colleagues might typically be to assert that the sample of 25 pupils was too small, and to reject the experimenter's conclusions despite his evidence. Zero correlation is generally not an appropriate null hypothesis for mental test data; it is rather an alternative hypothesis which might be acceptable in a given case only after weighty evidence against the general run of experience with positive correlations for such data.

It might be noted that the newer Lawley-Rao tests of significance might also find that the over-all rank of the observed correlations, using communalities, was not significantly different from zero for the same data above. The zero-rank hypothesis would still find few buyers among psychologists.

Statistical textbooks almost invariably start with zero as a null hypothesis for correlation coefficients. Historical reasons for this may lie in the early use of mathematical statistics in biology and other fields. Whatever the cause, such a habit need not be adopted uncritically by psychometricians, especially for mental test data with which they have had so much experience. It is just such experience which makes psychometricians regard not only zero rank, but also Spearman's and Thurstone's hypotheses of small rank to be alternatives, and to be acceptable only if the data warrant rejecting the hitherto more plausible null hypothesis of great complexity.

How Did the Shoe Get On the Other Foot?

Despite the fact that the Spearman-Thurstone type of hypothesis is relatively novel, some followers of Thurstone appear to have reversed the problem of how to test it. They seem to accept the hypothesis of small rank a priori, and demand substantial evidence before they will believe in large rank. This is indicated in their approach to the question of "when to stop factoring." That question arose historically out of certain computing routines aimed at extracting one factor at a time from an empirical correlation matrix. Somehow, the notion became current that there was a danger of extracting

too many factors, or more than were warranted by the data. This fear patently puts the shoe on the other foot: it makes small rank the null hypothesis, or it necessitates acceptance of the Thurstonian hypothesis unless there is strong evidence to the contrary. Such a lenient attitude towards a new hypothesis is quite unusual in the history of science.

But the attitude is not entirely unambiguous. Proponents of Thurstone's thesis recognize to a certain extent that it should not be accorded the status of a null hypothesis. They express concern about obtaining low rank as a mere algebraic artifact from use of communalities, a concern that often involves belief in inequality (1) above. Thus, it has been concluded from (1) that "Three tests can always be reduced to rank 1 in a finite number of ways $\cdots$ Four tests can always be reduced to rank 2 in an infinite number of ways$\cdots$ with 6 tests it is in general possible to attain rank 3 without any restrictions on the (observed correlation) coefficients$\cdots$" ([14], p. 92).

Belief in inequality (1) led Thurstone himself to suggest that the right member of (1) be the null hypothesis for the minimal rank [cf. 18]. However, no one has come forward with an explicit computing routine based on such a null hypothesis. Inequality (1) can hardly be taken as a rigorous point of departure, for it is in fact false. This is easy to show by a simple example, as in the following section. The best possible universal upper bound to minimal rank m is actually $n - 1$.

The Best Possible General Upper Bound: $n - 1$

It is well known that if R is a nonsingular correlation matrix of order n, then it can always be reduced at least to rank $n - 1$ with communalities [18]. One way of accomplishing such a reduction is to subtract the smallest latent root of R from each main diagonal element. If this root is of multiplicity p, then the rank of the modified R is $n - p$ [5]. Another way is to select one value of j, and replace the corresponding diagonal element of R by the square of the multiple correlation coefficient of the observed variable concerned on the $n - 1$ remaining variables; R modified this way must be exactly of rank $n - 1$ [10]. (There are n ways of doing this, since any one of the j diagonal elements may be thus modified to do the trick.)

Each of the above methods not only reduces the rank of R, but also leaves the modified R Gramian, as is required for real-valued factor scores and loadings. The resulting communalities also do not exceed unity, and hence are proper.

Granted that nonsingular R can always be reduced to rank $n - 1$, can one in general do better than this? Exhibiting a nontrivial example wherein it is impossible to go below rank $n - 1$ suffices to answer this question from a purely algebraic point of view.

Consider the following symmetric matrix R_z of order n:

$$(2) \qquad R_x = \begin{bmatrix} x_1 & r_1 & 0 & 0 & \cdots & 0 & 0 \\ r_1 & x_2 & r_2 & 0 & \cdots & 0 & 0 \\ 0 & r_2 & x_3 & r_3 & \cdots & 0 & 0 \\ 0 & 0 & r_3 & x_4 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \cdots & x_{n-1} & r_{n-1} \\ 0 & 0 & 0 & 0 & \cdots & r_{n-1} & x_n \end{bmatrix}.$$

Assume that the r_j are fixed numbers, observed correlation coefficients for the problem of factor analysis, none of which vanish,

$$(3) \qquad r_j \neq 0, \qquad (j = 1, 2, \cdots, n-1).$$

The main diagonal elements x_j $(j = 1, 2, \cdots, n)$ are arbitrary and may be determined so as to minimize the rank of R_x. All other elements of R_x vanish.

No matter what values are chosen for the x_j, the resulting rank of R_x cannot be less than $n - 1$. To prove this, it is sufficient to examine the submatrix of order $n - 1$ obtained by omitting the first row and last column of R_x. This is a triangular matrix, whose main diagonal elements are the r_j and in which all elements below the main diagonal vanish. Its determinant is merely the product of the $n - 1$ coefficients r_j, and is unaffected by the choice of the x_j. By virtue of (3), this determinant does not vanish, so R_x has at least one submatrix of rank $n = 1$. Hence, R_x itself can never have rank less than $n - 1$, no matter what values are used in the main diagonal. This establishes our theorem that $n - 1$ cannot be improved on as a universal upper bound to minimal rank.

Relative Algebraic Incidences of Large and Small Minimal Proper Ranks

Note that our example of R_x establishes a bound for rank for the set of all symmetric matrices, not just Gramian ones. It has been recognized by those who have believed in inequality (1) that the communalities implied need not be proper. Some communalities might be larger than the self-correlations of their corresponding observed variables, yielding so-called Heywood cases [17], and/or the modified R might not be Gramian. But our example shows that even with improper communalities, rank need not be reducible below $n - 1$. No restrictions whatsoever were put on the x_j, yet R_x cannot be of rank less than $n - 1$.

If the problem is restricted to use of proper communalities, it becomes even clearer that small rank cannot be the *general* algebraic case. In [10] it was shown that every nonsingular correlation matrix R of order n has a one-to-one correspondence with another nonsingular correlation matrix

R^* or order n (where R^* is directly related to R^{-1}) such that if m and m^* are the minimal ranks to which R and R^* are reducible respectively by use of proper communalities, then in general

$$m \geqq n - m^*. \tag{4}$$

Thus, m and m^* cannot in general simultaneously be small compared with n. If $m^* < n/2$, then $m \geqq n/2$ and if $m < n/2$ then $m^* \geqq n/2$.

Therefore, considering the set of all possible nonsingular correlation matrices of order n, the subset that can be properly reduced to rank $m < n/2$ cannot be larger (have a larger cardinal number) than the subset for which proper minimal rank is $m \geqq n/2$. More generally, the subset of matrices that are properly reducible to rank 1 cannot be larger than the subset of matrices that are properly reducible to rank $n - 1$; the subset for which $m = 2$ cannot be larger than the subset for which $m = n - 2$, etc.

Since m and m^* can be, and often are, *large simultaneously*, should any probability measure be attached to the set of all nonsingular correlation matrices, it would in general show that large minimal proper rank is more probable than small proper rank. (The measure implied here is over population matrices; similar reasoning will of course hold for sample matrices, but the latter case is not our present concern.)

Psychological Considerations

Some readers may accept the algebraic theorems of the last two sections, but remain unimpressed as to their psychological relevance. On the one hand, the algebra of the previous section may seem to them too general, and hence not necessarily appropriate to psychological data. On the other hand, the specific matrix R_s defined by (2) is hardly typical of observed mental test correlation matrices. Let us look into such objections.

There are some who believe that the Thurstonian approach is an appropriate one for any correlation matrix. Such students should welcome general algebraic theorems. Similarly, mathematical statisticians who are concerned with developing general statistical tests of significance or decision rules should want to consider such a general algebra.

To complain about excessive generality of the preceding section may be but another example of putting the shoe on the wrong foot. The onus of proof should be on the complainers: if they believe psychological data differ from others, it is they who should provide the evidence.

It so happens that the general algebraic case leads to the same type of null hypothesis as general psychological considerations: large rank. No actual contradiction exists between the general algebraic case and the more purely psychological one with regard to over-all orientation on the problem of rank. Psychologists cannot invoke general algebra nor can algebraists invoke psychology to make out an a priori case for small rank.

The Simplex As a Counter Example

That R_x is not typical of psychologically observed correlation matrices is certainly correct. There is, however, a large class of matrices approximated in psychological practice for which $n - 2$ is the minimal rank. These matrices are closely related to R_x^{-1}, and proof of their rank properties is simplified by use of the rank theorem on R_x .

Abundant empirical evidence testifies that when the kind of mental ability tested is held constant and only level of complexity is varied among tests, the resulting correlation matrix will tend to be some type of simplex [6, 7, 11]. Let R be the correlation matrix for a perfect additive simplex, which is one of the possible types. Thus, if r_{jk} is the coefficient in row j and column k of R,

$$r_{jk} = a_j/a_k \qquad (j \leqq k), \tag{5}$$

where a_j is some coefficient associated with the jth variable ($j = 1, 2, \cdots, n$). According to (5), tests of more similar levels of complexity (whose coefficients a_j are more nearly equal) will correlate more highly with each other than will tests of less similar levels of complexity. The subscripts j are assigned the tests according to their order of complexity.

It has been shown in [6] that if R is defined by (5) then R^{-1} is precisely of the form R_x in (2), or

$$R^{-1} = R_x. \tag{6}$$

To what rank can R be reduced by modifying its main diagonal? Let D be an arbitrary diagonal matrix of order n. What is the minimum rank possible for $R - D$? To answer this, first note that the rank of $R - D$ is the same as that of $D - R$. It then follows from the identity

$$R^{-1}(D - R) \equiv R^{-1}D - I, \tag{7}$$

where I is the unit matrix of order n, that the rank of $R - D$ is that of $R^{-1}D - I$; for the rank of the left member of (7) is that of $D - R$ since premultiplying by a nonsingular matrix does not affect rank. Recalling (6), it is established that the rank of $R - D$ equals that of $R_xD - I$.

While $R_xD - I$ is an asymmetric matrix, if written out explicitly as is R_x in (2), it always has a nonvanishing minor determinant of order $n - 2$. Two cases have to be considered: where D has no vanishing diagonal elements, and where D has one or more vanishing diagonal elements. Proof is left to the reader. It follows that $R_xD - I$ cannot be of rank less than $n - 2$ so neither can $R - D$, which was to be proved.

Notice that D has *not* been restricted here to yield proper communalities. The diagonal elements of D may be positive or negative, and $R - D$ need not be Gramian. Regardless, $R - D$ can never attain rank less than $n - 2$

when R is a nonsingular correlation matrix for a perfect additive simplex defined by (5).

Further Structural Psychological Possibilities

Artificiality remains in using the algebra of a perfect simplex such as defined by (5); data do not conform to this in practice, because in essence (5) represents the structure of the correlations after correction for uniqueness. Raw correlations conform more directly to the form

$$r_{jk} = a_j b_k \qquad (j < k), \tag{8}$$

where b_k is proportional to a_k^{-1}, so that the communality of test j is given by the right member of (8) when $j = k$: $h_j^2 = a_j b_j$. Given a raw matrix with off-diagonal elements defined by (8) and with unities down the main diagonal, one still cannot reduce its rank to less than $n - 2$ by modifying the main diagonal. This can easily be seen by the proportionality relation of (5) to (8); if one type of matrix can be reduced in rank by tampering with the main diagonal, so can the other. Proof is left to the reader. (Write $b_k = c/a_k$, where c is some constant, and the proof is almost immediate.)

The inverse matrix for such an imperfect simplex as (8) is *not* in general of the form R_z ; the nice pattern of exactly zero cells no longer holds. Regardless, rank cannot be reduced below $n - 2$ by communalities.

For the more general type of simplex analyzed in [7], it can be shown that rank cannot be reduced in general below $n - 3$, whether in the perfect case or in the imperfect case where the δ-law of deviation holds. There are many classes of psychological data for which such a model should be appropriate, since it is formally identical with the stochastic process of uncorrelated increments [3]. One psychological interpretation may be that of increments in complexity, as we have already indicated. Another interpretation may be that of increments in speed.

The Example of Speed; Facet Design

A study involving different speed levels of each of three different kinds of content recently was published by Lord [15]. Looking at the submatrices of correlation coefficients for each kind of content separately shows an obvious gradient typical for a simplex. Despite the avowed intention to study speed factors, Lord's design of the test battery does not make possible a clear analysis of the speed process for each kind of content separately, and hence not of the interrelations between the various contents at different speed levels. Instead, a relatively small number of factors were extracted, since it was believed that the estimated communalities had actually reduced the rank of the over-all matrix considerably; and attempts were made to interpret the factors.

The relative inadequacy of such an attempt at rank reduction should become clear by considering a sub-battery of test for one content alone, say

arithmetic. Suppose a pool of items of similar complexity were available, and tests were constructed from them to be varied only in the time limits of administration: 1 minute, 2 minutes, $\cdots$, n minutes. What should the structure of the resulting correlation matrix be? If the increments in scores between successive time limits are uncorrelated with each other—as Lord's data tend to show—then communalities cannot reduce rank below $n - 3$. Furthermore, even such rank reducing communalities would not be the most natural ones to use, for they would destroy the possibility of obtaining the parsimoniously structured inverse matrix that exists for a simplex [7], so important for prediction purposes.

Since speed is a pervasive concept in psychological abilities, as well as level of complexity, there is thus further psychological argument against expecting small reduced rank to be possible with actual psychological data.

Recent developments in facet design for kind of complexity, as distinguished from level of complexity and from speed, provide further examples of possible psychological structures for which an explicit algebraic structure or common-factor pattern may be specified [12]. Given such explicit algebra, it should be possible to develop theorems on what communalities can and cannot do for such cases. The rapidity with which such designs attain substantial complexity (but in a systematically simple manner) again suggests that the concept of communality should be dissociated from that of rank reduction. Instead, the concept might be defined in terms of the δ-law of deviation [6]: find that set of unique factor loadings to satisfy the δ-law such that the best fit can be obtained from the hypothesized algebraic structure (according to the facet design) to the observed correlation matrix. This is a type of parsimony that has little or nothing to do with rank.

It is not the purpose of the present paper to go into alternatives to the rank-reduction approach to the communality problem. Some have already been indicated, and others may be possible.

The Approximation Argument

Some who have read this paper to this point may still feel no great need for abandoning current computing procedures that attempt to stop at as small a rank as possible. They may grant all the above algebraic and psychological arguments, but believe that these are beside the point. That there are many common factors in most given cases they may concede, even as many as there are observed tests. But a factor analyst, they may claim, should be interested only in that small number that helps approximate the observed correlations, and ignore the remaining numerous common factors, which surely exist, that account for the discrepancies. That is, apart from unique factors, they advise ignoring all common factors that contribute but little to the intercorrelations. Since only approximations are sought, the exact theorems on rank are irrelevant, they might say.

In reply to such an argument, it might be pointed out that usual,

univariate concepts of approximation may break down when considering multivariate problems like that of factor analysis. What is a good approximation to observed data for the multivariate case? There is no simple answer to this, for there are many features of a multivariate distribution that one might be interested in estimating; a good approximation to one aspect may lead to a very bad approximation to another aspect.

For example, if R is any observed correlation matrix, it is known that the best single set of factor loadings for approximating the elements of R is the latent vector associated with the largest latent root of R. By the same token, the best single set of components for estimating the elements of R^{-1} is the latent vector associated with the smallest latent root of R: what is best for R is worst for R^{-1}, and vice versa. Why should the approximation problem be addressed to R, as is done traditionally, rather than to R^{-1}? It is R^{-1} that is directly involved in prediction problems concerning the observed variables, and from this point of view R^{-1} is more important than R. Furthermore, if D is nonsingular, it is easy to see that the rank of $R - D$ is that of $R^{-1} - D^{-1}$ (postmultiply (7) through by D^{-1}, letting R now be arbitrary), but the approximation problem takes on a different aspect in each case if posed in terms of rank.

The ambiguity as to what ought to be approximated no doubt has contributed some confusion to the communality problem. (It might be noted that in the radex approach, the approximations to both R and R^{-1} are checked.) Consideration of sizes of latent roots alone is not sufficient for justifying approximations.

If one frowns on latent root and principal axis solutions generally for factor analysis, but believes rotations of axes should be made to be meaningful, one cannot be in a good position even to raise the problem of approximation. There is no general way to array the common factors in order of their *importance*. To the contrary, rotations can always be made to introduce substantial loadings on as many factors as tests. Argument for small rank in terms of approximation still seems to put the shoe on the wrong foot.

The Fallacy Behind Inequality (1)

Inequality (1) was arrived at by similar lines of reasoning (although different in detail) by Thurstone [18] and Ledermann [14]. Burt [1] has pointed out an earlier treatment by Shepard. Each author essentially proceeds to set up equations which must be satisfied if R is to be reducible to rank m by modifying its main diagonal elements. They count two things: the number of equations and the number of unknowns. Then they hypothesize that the equations are solvable if the number of unknowns is not less than the number of equations, and arrive at (1).

None of the authors cited apparently claims (1) as an actual theorem. But they, and many students following them, act as though it were a theorem

for all practical purposes. The fallacy, of course, results from the fact that equations need not be solvable merely because the number of unknowns equals, or even exceeds, the number of equations.

Many students have fallen into the habit of believing that n equations in n unknowns are in general solvable. High school and college drill on *linear* equations of course should lead to such a belief, since almost all the examples the teachers choose are indeed solvable. It is another matter to inquire into the natural frequency of solvability. Even if one should agree, on geometric grounds, that solvability should hold more often than not for *linear* equations (i.e., collinearity or linear dependence is less likely to occur than not), this says little or nothing about nonlinear equations. For example, consider two simultaneous equations in two unknowns whose loci are each ellipses: are arbitrary ellipses in a plane more likely to intersect than not? Merely counting equations and unknowns is not very helpful here.

The equations behind proposed inequality (1) are of quite a complex curvilinear nature. It borders on wishful thinking to believe a priori that they should be solvable at all in any given case, even if imaginary or other improper solutions be allowed. That it has been possible to develop concrete algebraic theorems on this problem of frequency of solvability, as discussed in this paper, may help dispel some prevalent miscomprehensions concerning communalities and rank.

What has happened to factor analysis in this regard may be similar to what happened much earlier with regard to belief in the normal (Gaussian) curve for psychological data. As someone once pointed out, psychologists thought it was the mathematicians who proved that the normal distribution must hold for their data, while mathematicians thought that use of its equation was justified because psychologists found it to hold empirically.

Some psychologists may have thought that making small rank the null hypothesis was justified by *algebraic* considerations, while mathematical statisticians may have thought that starting with such a null hypothesis was justified by *psychological* considerations. Neither case is correct. Algebra and psychology both indicate large rank to be the more proper null hypothesis for the communality problem for mental test data. If the null hypothesis is sustained, this implies other concepts of parsimonious structure are called for than that of small rank.

REFERENCES

[1] Burt, C. Bipolar factors as a cause of cyclic overlap. *Brit. J. Psychol., Statist. Sect.*, 1952, **5**, 197–202.

[2] Cattell, R. B. Index universel pour les facteurs psychologiques. In *L'analyse factorielle et ses application*, Paris: Centre National de la Recherche Scientifique, 1956.

[3] Doob, J. W. *Stochastic processes.* New York: Wiley, 1954.

[4] Guttman, L. Image theory for the structure of quantitative variates. *Psychometrika*, 1953, **18**, 277–296.

[5] Guttman, L. Some necessary conditions for common-factor analysis. *Psychometrika*, 1954, **19**, 149–161.
[6] Guttman, L. A new approach to factor analysis: the radex. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences*. Glencoe, Ill.: Free Press, 1954.
[7] Guttman, L. A generalized simplex for factor analysis. *Psychometrika*, 1955, **20**, 173–192.
[8] Guttman, L. The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. *Brit. J. statist. Psychol.*, 1955, **8**, 65–81.
[9] Guttman, L. Une solution au problème des communautés. *Bulletin du Centre d'Études et Recherches Psychotechniques*, 1956, **6**, 123–128.
[10] Guttman, L. "Best possible" systematic estimates of communalities. *Psychometrika*, 1956, **21**, 273–285.
[11] Guttman, L. Empirical verification of the radex structure of mental abilities and personality traits. *Educ. psychol. Measmt*, 1957, **17**, 391–407.
[12] Guttman, L. What lies ahead for factor analysis? *Educ. psychol. Measmt*, in press.
[13] Lawley, D. N. The estimation of factor loadings by the method of maximum likelihood. *Proc. Roy. Soc. Edin.*, 1940, **60**, 64–82.
[14] Ledermann, W. On the rank of the reduced correlation matrix in multiple-factor analysis. *Psychometrika*, 1937, **2**, 85–93.
[15] Lord, F. M. A study of speed factors in tests and academic grades. *Psychometrika*, 1956, **21**, 31–50.
[16] Rao, C. R. Estimation and tests of significance in factor analysis. *Psychometrika*, 1955, **20**, 93–111.
[17] Thomson, G. H. *The factorial analysis of human ability*. (5th ed.) London: Univ. London Press, 1951.
[18] Thurstone, L. L. *Multiple-factor analysis*. Chicago: Univ. Chicago Press, 1947.
[19] Wrigley, C. The distinction between common and specific variance in factor theory. *Brit. J. statist. Psychol.*, 1957, **10**, 81–98.

Manuscript received 7/19/57
Revised manuscript received 1/3/58

PSYCHOMETRIKA—VOL. 23, NO. 4
DECEMBER, 1958

A MARKOV MODEL FOR DISCRIMINATION LEARNING*

RICHARD C. ATKINSON
UNIVERSITY OF CALIFORNIA, LOS ANGELES

A theory for discrimination learning which incorporates the concept of an observing response is presented. The theory is developed in detail for experimental procedures in which two stimuli are employed and two responses are available to the subject. Applications of the model to cases involving probabilistic and nonprobabilistic schedules of reinforcement are considered; some predictions are derived and compared with experimental results.

This paper is a preliminary attempt to develop a quantitative theory of discrimination learning. For simplicity, the discussion will be limited to two-response problems, but the formulation can be extended readily to certain n-response situations. The model corresponds in some respects to theoretical analysis of discrimination learning presented by Burke and Estes [5] and Bush and Mosteller [6]. In particular, the stimulus conceptualization and response conditioning process are similar to their formulations. The model, however, differs from their work in that an orienting or observing response [16] is postulated. This additional feature leads to predictions which, for some experimental parameter values, are markedly different from those made by either Burke and Estes or Bush and Mosteller, while for other parameter values the predictions are identical. Interrelations among these models will be considered later.

The theory is designed to analyze behavior in an experimental setup where two stimuli, designated T_1 and T_2, are employed and two responses, A_1 and A_2, are available to the subject. Each trial begins with the presentation of either T_1 or T_2; the probability of T_1 is β, and the probability of T_2 is $1 - \beta$. Following T_1, an A_1 response is correct with probability π_1, and an A_2 response is correct with probability $1 - \pi_1$. Following T_2, an A_1 response is correct with probability π_2, and an A_2 response is correct with probability $1 - \pi_2$.

The traditional type of discrimination problem is described when $\pi_1 = 1$ and $\pi_2 = 0$. The subject must learn to respond with A_1 to the presentation of T_1 and respond with A_2 to the presentation of T_2. A form of discrimination learning, involving probabilistic schedules of reinforcement, is specified when π_1, $\pi_2 \neq 0$ or 1. This type of discrimination problem has been only recently investigated [9, 10, 14].

*This research was supported by a grant from the National Science Foundation.

Theoretical Concepts

Stimulus Representation

The stimuli T_1 and T_2 are to be represented conceptually as two sets of stimulus elements, which are designated $\mathfrak{S}_1$ and $\mathfrak{S}_2$, respectively. Further, a set C is designated which represents those stimulus elements common to sets $\mathfrak{S}_1$ and $\mathfrak{S}_2$ ($C = \mathfrak{S}_1 \cap \mathfrak{S}_2$), i.e., those stimulus events common to the presentation of either T_1 or T_2. In regard to the size of the C set, an *index of similarity* between T_1 and T_2 can be defined; the larger the relative size of C with respect to $\mathfrak{S}_1$ and $\mathfrak{S}_2$ the greater the similarity between the stimuli [6].

To simplify subsequent notation let the set S_1 be all stimulus elements in $\mathfrak{S}_1$ but not in C. Similarly, S_2 is the set of all stimulus elements in $\mathfrak{S}_2$ but not in C. Specifically,

$$S_1 = \mathfrak{S}_1 - C, \qquad S_2 = \mathfrak{S}_2 - C. \tag{1}$$

Let N_1, N_2, and N_c be the number of elements in sets S_1, S_2, and C, respectively. Finally, define

$$W_1 = \frac{N_1}{N_1 + N_c}, \qquad W_2 = \frac{N_2}{N_2 + N_c}. \tag{2}$$

Orienting Response

It is hypothesized that the organism makes one of two responses at the start of each trial, either an orienting response or a nonorienting response. These responses are designated O and $\overline{O}$, respectively. If O occurs, then the organism is exposed to the unique stimulus elements on that trial. More specifically, the organism will be exposed to only the S_1 stimulus elements if T_1 is presented and to only the S_2 stimulus elements if T_2 is presented. If, on the other hand, $\overline{O}$ occurs, then the organism is exposed to both unique and common stimulus elements on the trial. The organism will be exposed to both S_1 and C stimulus elements if T_1 is presented and to both S_2 and C stimulus elements if T_2 is presented.

It is assumed that the O and $\overline{O}$ responses are elicited by a set $\mathfrak{O}$ of stimuli associated with the beginning of the trial. Thus, the sequence of events on a given trial is as follows.

(i) The onset of the trial is associated with the presentation of a set of stimulus elements $\mathfrak{O}$.

(ii) $\mathfrak{O}$ elicits either an O or $\overline{O}$ response.

(iii) If O occurs, the organism is exposed to the S_1 set on T_1 trials and to the S_2 set on T_2 trials. If $\overline{O}$ occurs, the organism is exposed to S_1 and C on T_1 trials and to S_2 and C on T_2 trials.

Conditioning Relations and Response Probability

On any trial of an experiment, all elements of a given stimulus set are

conditioned to one and only one response. The entire $\mathfrak{S}$ set is conditioned to either O or $\overline{O}$. Similarly, the S_i set ($i = 1$ or 2) is conditioned to either A_1 or A_2 and the C set is conditioned to either A_1 or A_2.

The probability of a response in the presence of particular stimulus elements is defined as the proportion of stimulus elements conditioned to the response [1, 8]. Thus, the probability of O on trial n, $p_n(O)$, is either 1 or 0, depending on whether the $\mathfrak{S}$ set is conditioned to O or $\overline{O}$ at the start of trial n. The probability of A_1 when only S_i ($i = 1$ or 2) is presented (i.e., when an O response has occurred at the start of the trial) is 1 or 0 depending on whether the S_i set is conditioned to A_1 or A_2. Finally, the probability of A_1 when S_i and C are presented (i.e., when an $\overline{O}$ has occurred at the start of the trial) is, (i) 1 if both the S_i set and the C set are conditioned to A_1, (ii) W_i if the S_i set is conditioned to A_1 and the C set is conditioned to A_2, (iii) $1 - W_i$ if the S_i set is conditioned to A_2 and the C set is conditioned to A_1 and, (iv) 0 if both the S_i set and the C set are conditioned to A_2.

Conditioning Process

A single parameter θ is assumed which governs the conditioning of stimulus sets. On a given trial all elements from $\mathfrak{S}$ and *available* elements from $\mathfrak{S}_i$ will be conditioned with probability θ to an appropriate response, and the conditioned status of all elements will remain unchanged with probability $1 - \theta$. Only those elements in $\mathfrak{S}_1$ and $\mathfrak{S}_2$ which are exposed to the organism on a given trial are available for conditioning. If an O response is made at the start of the trial, then either S_1 or S_2 is available for conditioning on the trial. If an $\overline{O}$ response is made, then either S_1 and C or S_2 and C are available for conditioning. Specifically the following cases encompass all possibilities.

(1) $O \rightarrow T_i \rightarrow A_j \rightarrow$ *correct.* An observing response is made and makes set S_i available. The set S_i elicits A_j, which is designated correct. Given this sequence of events, there is (i) a probability θ that all elements in $\mathfrak{S}$ will be conditioned to O and all elements in S_i will be conditioned to A_j and, (ii) a probability $1 - \theta$ that the conditional status of the element will remain unchanged.

(2) $O \rightarrow T_i \rightarrow A_j \rightarrow$ *not correct.* An observing response is made, and makes set S_i available. The set S_i elicits A_j, which is incorrect. Given this sequence of events there is (i) a probability θ that all elements in $\mathfrak{S}$ will be conditioned to $\overline{O}$ and all elements in S_i will be conditioned to A_j other than the one which occurred on the trial, and (ii) a probability $1 - \theta$ that the conditional status of the elements will remain unchanged.

(3) $\overline{O} \rightarrow T_i \rightarrow A_j \rightarrow$ *correct.* A nonobserving response is made and makes sets S_i and C available; the sets S_i and C elicit A_j, which is correct. Given this sequence of events there is (i) a probability θ that all elements in $\mathfrak{S}$ will be conditioned to $\overline{O}$ response and all elements in both S_i and C will be condi-

tioned to A_j , and (ii) a probability $1 - \theta$ that the conditional status of the elements will remain unchanged.

(4) $\overline{O} \rightarrow T_i \rightarrow A_j \rightarrow$ *not correct.* A nonobserving response is made and makes sets S_i and C available; the sets S_i and C elicit A_j , which is incorrect. Given this sequence there is (i) a probability θ that all elements in $\mathfrak{O}$ will be conditioned to O and all elements in both S_i and C will be conditioned to A_j other than the one which occurred on the trial, and (ii) a probability $1 - \theta$ that the conditional status of the elements will remain unchanged.

The above assumptions for conditioning and response probability are different from those postulated by Estes and Burke in their stimulus sampling model [5, 8]. No attempt will be made to compare the two sets of assumptions, but it should be noted that the ideas fundamental to the model presented in this paper initially were formalized within the framework of the Estes and Burke stimulus sampling theory. Unfortunately, the mathematical analysis resulted in a system of difference equations for which methods of solution are not known. Consequently, certain simplifying assumptions were made which yielded the present model. The difference in complexity between the stimulus sampling formulation and the present analysis is reflected in the state spaces of the respective stochastic processes. For the model presented in this paper the state space includes only six points $\{0, W_1 , W_2 , 1 - W_1 , 1 - W_2 , 1\}$, while the state space for the stimulus sampling model is defined on the closed interval [0, 1].

Mathematical Formulation

Given this conditioning process and the assumption that all stimulus elements in a particular set ($\mathfrak{O}$, S_1 , C, S_2) are conditioned to the same response at the start of the first trial, an organism can be described as being in one of sixteen possible states on any trial. A state will be specified by an ordered four-tuple where:

(i) the first member of the tuple indicates whether all elements in set $\mathfrak{O}$ are conditioned to O or $\overline{O}$;
(ii) the second member indicates whether elements in S_1 are conditioned to A_1 or A_2 ;
(iii) the third member indicates whether elements in C are conditioned to A_1 or A_2 ;
(iv) the fourth member indicates whether elements in S_2 are conditioned to A_1 or A_2 .

As an example, the state $\langle \overline{0}, 1, 1, 2 \rangle$ indicates that the $\mathfrak{O}$ set is conditioned to $\overline{O}$, S_1 is conditioned to A_1 , C is conditioned to A_1 , and S_2 is conditioned to A_2 . If the organism is in this state at the start of trial n, then if T_1 is presented an A_1 will occur, and if T_2 is presented an A_2 will occur with

probability W_2 and an A_1 with probability $1-W_2$. The states will be assigned identifying numbers as follows.

1. $\langle 0, 1, 1, 1\rangle$	5. $\langle 0, 2, 1, 1\rangle$	9. $\langle \bar{0}, 1, 1, 1\rangle$	13. $\langle \bar{0}, 2, 1, 1\rangle$
2. $\langle 0, 1, 1, 2\rangle$	6. $\langle 0, 2, 1, 2\rangle$	10. $\langle \bar{0}, 1, 1, 2\rangle$	14. $\langle \bar{0}, 2, 1, 2\rangle$
3. $\langle 0, 1, 2, 1\rangle$	7. $\langle 0, 2, 2, 1\rangle$	11. $\langle \bar{0}, 1, 2, 1\rangle$	15. $\langle \bar{0}, 2, 2, 1\rangle$
4. $\langle 0, 1, 2, 2\rangle$	8. $\langle 0, 2, 2, 2\rangle$	12. $\langle \bar{0}, 1, 2, 2\rangle$	16. $\langle \bar{0}, 2, 2, 2\rangle$

For these conditioning assumptions and the experimental parameters β, π_1, and π_2 a transition matrix P describing the learning process can be derived and is presented in Table 1. To simplify notation, in writing the P matrix let $a = \theta\beta\pi_1$, $b = \theta\beta(1-\pi_1)$, $c = \theta(1-\beta)\pi_2$, and $d = \theta(1-\beta)(1-\pi_2)$.

The state at the start of trial n is listed on the row, and the state at the start of trial $n+1$ is listed on the column. For example, a, the entry in row 15, column 1, is the conditional probability of being in state $\langle 0, 1, 1, 1\rangle$ at the start of trial $n+1$ given that the organism was in state $\langle \bar{0}, 2, 2, 1\rangle$ at the start of trial n.

Let $u_i(n)$ be the expected probability of being in state i ($i = 1$ to 16) at the start of trial n, where the first experimental trial is $n = 0$. Define the row matrix

$$U(n) = [u_1(n), u_2(n), \cdots, u_{16}(n)]. \tag{3}$$

Further, let P represent the one-stage transition matrix of order sixteen presented above, where p_{ij} is the conditional probability of being in state j on trial $n+1$, given that the system was in state i on trial n. Then the Markov process describing discrimination learning at the start of trial n is

$$\vec{U}(n) = \vec{U}(0)P^n. \tag{4}$$

(For a general consideration of finite Markov processes see [4, 11, or 12]. For applications of Markov processes to learning see [2, 3, 13].)

Experimentally it is impossible to identify individual states of the process on a particular trial. That is, given information about the type of trial and the A_i response which occurred, what state the organism was in at the start of the trial cannot be specified. For example, if T_1 is presented and A_2 occurs, which of the sixteen states the organism was in when the A_2 occurred cannot be established unequivocally. In fact, for this particular combination, any one of the following eight states would have been possible: 1, 2, 3, 4, 9, 10, 11, or 12. Obviously, this confounding is due to the fact that O and $\bar{O}$ responses have been postulated which are not observable.

Since trial descriptions and theoretical states cannot be placed in one-

TABLE 1

The Transition Matrix P
(Zero terms are indicated by blank cells.)

States	States															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	$1-\theta+a+c$									d			b			
2		$1-\theta+a+d$							c					b		
3			$1-\theta+a+c$									d			b	
4				$1-\theta+a+d$							c					b
5					$1-\theta+b+c$				a					d		
6						$1-\theta+b+d$				a			c			
7							$1-\theta+b+c$				a					d
8								$1-\theta+b+d$				a			c	
9				d			b		$1-\theta+a+c$							
10	W_2c			$(1-W_2)d$				b	$(1-W_2)c$	$1-\theta+a$		W_2d				
11	$(1-W_1)a+$ $(1-W_2)c$			W_2d			W_1b		W_1a+W_2c		$1-\theta$	$(1-W_2)d$			$(1-W_1)b$	
12	c	$(1-W_1)a$						W_1b		W_1a		$1-\theta+d$				$(1-W_1)b$
13	W_1a						$(1-W_1)b$	d	$(1-W_1)a$				$1-\theta+c$		W_1b	
14		W_1a			W_2c			$(1-W_1)b+$ $(1-W_2)d$		$(1-W_1)a$			$(1-W_2)c$	$1-\theta$		W_1b+W_2d
15	a				$(1-W_2)c$			W_2d					W_2c		$1-\theta+b$	$(1-W_2)d$
16		a			c											$1-\theta+b+d$

to-one correspondence, it is necessary (for an experimental evaluation of the theory) to define probabilities of events that are observable. Consequently, the following probabilities are of particular interest: $p_n(A_1 \mid T_1)$, the expected conditional probability on trial n of A_1 given T_1 ; and $p_n(A_1 \mid T_2)$, the expected conditional probability on trial n of A_1 given T_2 . By inspection of the theoretical states it follows that

$$p_n(A_1 \mid T_1) = u_1(n) + u_2(n) + u_3(n) + u_4(n) + u_9(n) + u_{10}(n) + W_1[u_{11}(n) + u_{12}(n)] + (1 - W_1)[u_{13}(n) + u_{14}(n)], \tag{5}$$

and

$$p_n(A_1 \mid T_2) = u_1(n) + u_3(n) + u_5(n) + u_7(n) + u_9(n) + u_{13}(n) + W_2[u_{11}(n) + u_{15}(n)] + (1 - W_2)[u_{10}(n) + u_{14}(n)]. \tag{6}$$

Also, for analytical purposes, the probability of an observing response at the start of trial n will be useful.

$$p_n(O) = u_1(n) + u_2(n) + \cdots + u_8(n). \tag{7}$$

Analysis of the Model and Some Special Cases

To illustrate certain aspects of the theory, without going into extensive mathematical detail, several special cases will be considered. These cases are of particular interest experimentally. For each case illustrative learning functions will be presented. The computations have been performed with the following restrictions on parameter values: $W_1 = W_2 = W$; the initial probability of O was taken to be zero and the initial probability of A_1 to S_1 , S_2 , or C to be .5. That is, $u_1(0) = u_2(0) = \cdots = u_8(0) = 0$ and $u_9(0) = u_{10}(0) = \cdots = u_{16}(0) = 1/8$.

The computations were performed at the Western Data Processing Center on a 650 IBM computer. The program or punch program deck is available to anyone who is interested in generating theoretical functions for cases or parameter values not presented in this paper. The program is arranged so that the following information must be read into the computer memory: β, π_1 π_2 , θ, W_1 , W_2 , and the vector $U(0)$. The program will compute $U(n)$, $p_n(A_1 \mid T_1)$, $p_n(A_1 \mid T_2)$ and $p_n(O)$ for successive values of n and also asymptotic results for each of these quantities.

However, before examining special cases, a general result can be immediately established; the Cesàro asymptotic probability of any state in the process is independent of the value of θ when $\theta > 0$. This follows from the fact that the main diagonal of the matrix P has terms of the form $(1 - \theta) + \theta X$, while all other nonzero terms are of the form θY. For the case where $\theta = 0$, $u_i(n) = u_i(0)$ for all n.

Traditional Discrimination Learning

The case in which $\pi_1 = 1$, $\pi_2 = 0$, and $\beta \neq 0$ or 1 describes the often investigated situation in which the subject is required to respond with A_1 to the presentation of T_1 and with A_2 to the presentation of T_2. An inspection of the P matrix indicates that, for these particular parameter values, the process eventually will be absorbed in either state $\langle 0, 1, 2, 2 \rangle$ or state $\langle 0, 1, 1, 2 \rangle$ and, therefore, asymptotically

$$
\begin{aligned}
p_n(O) &\xrightarrow{n} 1 \\
p_n(A_1 \mid T_1) &\xrightarrow{n} 1 \\
p_n(A_1 \mid T_2) &\xrightarrow{n} 0.
\end{aligned} \tag{8}
$$

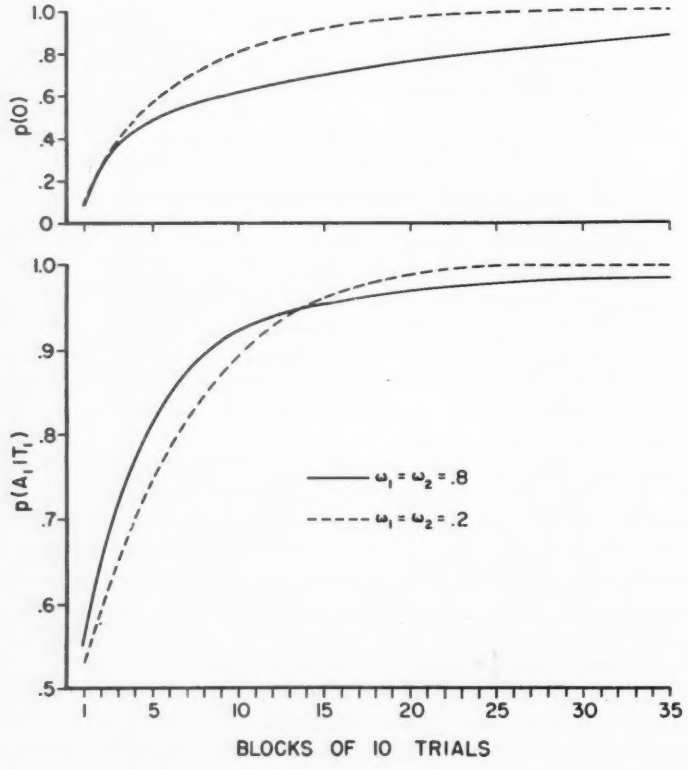

FIGURE 1

Theoretical discrimination learning function for two conditions of stimulus similarity. The case in which $\pi_1 = 1$, $\pi_2 = 0$, and $\beta = 1/2$.

Figure 1 presents several theoretical curves of $p_n(A_1 \mid T_1)$ in blocks of ten trials when $\theta = .05$ and $\beta = .5$. The curves for $p_n(A_1 \mid T_2)$ are not presented, since for $\beta = .5$ and the above initial conditions $p_n(A_1 \mid T_2) = 1 - p_n(A_1 \mid T_1)$. An interesting result is the relation between the functions for different values of W. Taking $W = .8$ as the comparison function, on early trials the $W = .2$ curve is below the comparison curve. However, by approximately trial 130 the $W = .2$ curve crosses the comparison curve and remains above it as they both approach unity. This appears to be a general relationship for a fixed value of θ and the above set of initial conditions; given $W^* > W^{**}$ on early trials, the W^* curve for $p_n(A_1 \mid T_1)$ will fall above the W^{**} curve, but at some trial a crossover will occur, and thereafter the W^{**} curve will be above the W^* curve as both approach unity. A proof of this result has not been obtained; however, calculations using many different values of W and θ in no case established a counter example. This is an interesting prediction, and one which should be verifiable. Unfortunately, no evidence has been found in the literature to confirm or negate this result. Research is now under way on this problem and is designed to manipulate W experimentally by varying the apparent similarity between discriminanda.

The Estes and Burke Study

The case in which $\pi_1 = 1.0$, $\pi_2 = .5$, and $\beta = .5$ describes a form of discrimination learning investigated by Estes and Burke [9]. There are several aspects to the study, but for the present analysis only the acquisition process for the constant group will be considered.

Facing the subject is a circular array of 12 lights, and either the onset of the six lights on the left half of the panel or the six on the right half are designated as a T_1 trial and the onset of the other six as a T_2 trial. On each trial the subject makes either an A_1 or A_2 response; this is followed by a signal which informs the subject which response was correct.

In Figure 2 the observed conditional probabilities of $p(A_1 \mid T_1)$ and $p(A_1 \mid T_2)$ in blocks of 20 trials are presented. On the average, for a block of 20 trials there will be 10 T_1 trials and 10 T_2 trials. Consequently, in a block of 20 trials the observed value of $p(A_1 \mid T_1)$ for a given subject is based on approximately 10 observations and $p(A_1 \mid T_2)$ is also based on approximately 10 observations.

Listed on the same graph are some theoretical curves computed for $\theta = .05$ and for $W = .1$, .6, and .9. As can be seen, the $W = .1$ curves provide a fairly close fit to the observed values for $p_n(A_1 \mid T_1)$ and $p_n(A_1 \mid T_2)$. For this value of W the $p_n(A_1 \mid T_1)$ curve approaches an asymptotic level of .907, which closely approximates the observed terminal values. More interesting, however, is the theoretical curve of $p(A_1 \mid T_2)$ for $W = .1$. It starts out at .5, rises to a maximum value at approximately trial 40, and then monotonically decreases to an asymptotic level of .525. This initial

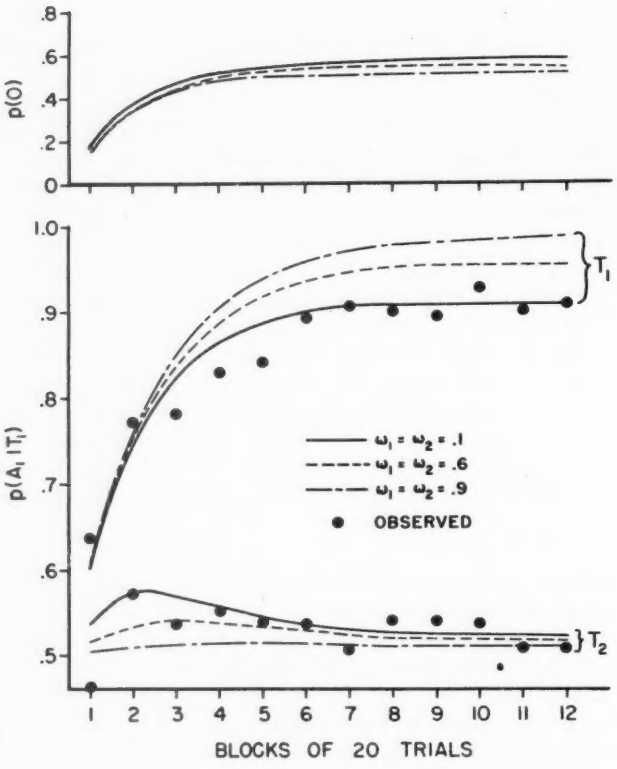

FIGURE 2

Observed values and theoretical functions of discrimination learning for the case in which $\pi_1 = 1$, $\pi_2 = 1/2$, and $\beta = 1/2$. Theoretical results for three conditions of stimulus similarity.

increase and subsequent decrease in the $p(A_1 \mid T_2)$ curve is evident in the Estes and Burke data and is an observation they emphasize in their discussion of the results.

A comparison of the curves in Figure 2 illustrates some general theoretical results for this special case; namely, $p_\infty(A_1 \mid T_1)$ is closer to unity and $p_\infty(A_1 \mid T_2)$ is closer to .5 the greater the value of W. Another prediction of experimental interest is that the smaller the value of W the greater the maximum value of $p_n(A_1 \mid T_2)$, and also the earlier the maximum will be reached. To illustrate, for the above computations, the maximum of $p_n(A_1 \mid T_2)$ for $W = .1$ was .575 and occurred on about trial 40, while the maximum for $W = .9$ was .540 and occurred on about trial 60.

The Popper and Atkinson Study

The final study to be considered used five groups [14]. For all groups $\pi_1 = .85$ and $\beta = .50$. The groups differed with respect to the π_2 parameter which took on the values .85, .70, .50, .30, and .15 for Groups I to V, respectively. In Figure 3 the observed proportions of A_1 responses following both T_1 and T_2 stimuli for the last 120 trials are presented. The experiment was run for a total of 320 trials. An inspection of the response curves by trials indicated that a stable level of responding had been reached during the last 120 trials. Consequently, the proportions presented in Figure 3 can be used as estimates of $p_\infty(A_1 \mid T_1)$ and $p_\infty(A_1 \mid T_2)$.

In fitting this data, the observed asymptotic value of $p_\infty(A_1 \mid T_1)$ for Group IV was used to evaluate W. The resulting estimate was $W = .1$. Using this value, predictions were then generated of $p_\infty(A_1 \mid T_1)$ for the other four groups and of $p_\infty(A_1 \mid T_2)$ for all five groups.

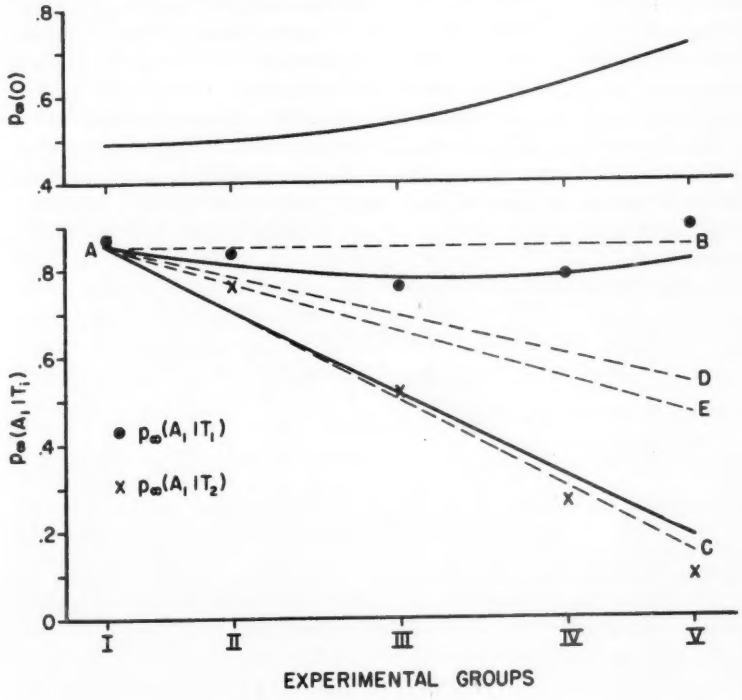

FIGURE 3

Predicted and observed asymptotic values for discrimination problems involving five probabilistic schedules of reinforcement.

The predicted values are given in Figure 3. No attempt will be made to present a detailed analysis of the data except to note that the general trends are approximated by the model. In particular, the model predicts a convex relation between $p_\infty(A_1 \mid T_1)$ and the value of π_2 which is reflected in the data. That is, for a fixed value of π_1 , $p_\infty(A_1 \mid T_1)$ first decreases and then increases as π_2 goes from .85 to .15. On the other hand, the theoretical values of $p_\infty(A_1 \mid T_2)$ for $W = .1$ are close to the π_2 value for all groups.

Discussion

No rigorous attempt has been made to test the model for the special cases considered. Nevertheless, qualitatively it appears that the theory accounts for some aspects of traditional types of discrimination learning and can be extended without modification to discrimination problems involving probabilistic reinforcement schedules.

Several studies currently in progress are designed specifically to test various features of the theory. The variables analyzed encompass a broad range of reinforcement schedules and include procedures designed to manipulate the index of similarity between stimuli. It is hoped that these investigations will provide a quantitative evaluation of the present theory and will lead to a more satisfactory formalization of discrimination learning.

For readers familiar with the theoretical work of Burke and Estes [5] and Bush and Mosteller [6] for discrimination learning, it may be helpful to examine some relations between their models and the one presented in this paper. Of particular interest are the asymptotic predictions generated by each model.

In the Bush and Mosteller model,

$$p_n(A_1 \mid T_1) \xrightarrow{n} \pi_1 \,, \qquad p_n(A_1 \mid T_2) \xrightarrow{n} \pi_2 \,, \tag{9}$$

independent of the value of β.

For Burke and Estes

$$p_n(A_1 \mid T_1) \xrightarrow{n} \pi_1 W_1 + (1 - W_1)\pi_e \,, \qquad p_n(A_1 \mid T_2) \xrightarrow{n} \pi_2 W_2 + (1 - W_2)\pi_e \,, \tag{10}$$

where $\pi_e = \beta\pi_1 + (1 - \beta)\pi_2$.

For the model presented in this paper, predicted asymptotes always lie between the predictions of Bush-Mosteller and Burke-Estes. That is, $p_\infty(A_1 \mid T_1)$ is bounded between π_1 and $\pi_1 W_1 + (1 - W_1)\pi_e$, while $p_\infty(A_1 \mid T_2)$ is bounded between π_2 and $\pi_2 W_2 + (1 - W_2)\pi_e$.

These relationships are illustrated in Figure 3. Equation (9) predicts that $p_\infty(A_1 \mid T_1)$ will fall on the straight line AB and $p_\infty(A_1 \mid T_2)$ will fall on the straight line AC. In contrast, (10) predicts for $W_1 = W_2 = .1$ that

$p_\infty(A_1 \mid T_1)$ will fall on the straight line AD, while $p_\infty(A_1 \mid T_2)$ will fall on the straight line AE. As indicated earlier, for the present model the predicted value of $p_\infty(A_1 \mid T_1)$ falls on the convex function bounded between the straight lines AB and AD, while $p_\infty(A_1 \mid T_2)$ falls on the function bounded between lines AC and AE. Actually to compute predictions for the Burke-Estes model, one would estimate W from the data of one group, using an estimation procedure appropriate to their model, and then predict the results of the other groups, as was done for the model presented in this paper. The results undoubtedly would be different from those indicated by lines AD and AE in Figure 3, but would fall on straight lines with origins at A.

In conclusion, it appears that this model generates some interesting predictions regarding both reinforcement schedules and similarity between discriminanda. Objections might be raised concerning the particular assumptions that were selected, but in the final analysis their evaluation will be determined in the laboratory. Nevertheless, several aspects of the theory leave the author uneasy; one feature is particularly disturbing and deserves comment. Reference is made to the assumption in which a single conditioning parameter θ is postulated. In essence, this assumption requires that the acquisition of O or $\overline{O}$ will progress at the same rate as the acquisition of A_1 or A_2 . Intuitively this seems an improbable state of affairs which may be approximated only for restricted experimental procedures. If this is the case, a change in the theory will be required such that two conditioning parameters are postulated, one governing the acquisition of O or $\overline{O}$ and the other A_1 or A_2 . This modification will still allow formalization of the model as a sixteen-state Markov process. However, the P matrix would have many more nonzero entries, and the theory would no longer yield asymptotic response probabilities which are independent of the conditioning parameters. These complications are not unmanageable, and if the modification proves necessary, theoretical predictions still can be generated easily.

REFERENCES

[1] Atkinson, R. C. A stochastic model for rote serial learning. *Psychometrika*, 1957, **22**, 87–96.

[2] Atkinson, R. C. and Suppes, P. An analysis of a two-person game situation in terms of statistical learning theory. *J. exp. Psychol.*, 1958, **55**, 369–378.

[3] Atkinson, R. C. and Suppes, P. An analysis of non-zero sum games in terms of a Markov model for learning. Tech. Rep. No. 9, Contract NR 171–034, Appl. Math. and Statist. Lab., Stanford Univ., 1957.

[4] Bartlett, M. S. *An introduction to stochastic processes.* Cambridge: Cambridge Univ. Press, 1955.

[5] Burke, C. J. and Estes, W. K. A component model for stimulus variables in discrimination learning. *Psychometrika*, 1957, **22**, 133–145.

[6] Bush, R. R. and Mosteller, F. A model for stimulus generalization and discrimination. *Psychol. Rev.*, 1951, **58**, 413–423.

[7] Bush, R. R. and Mosteller, F. *Stochastic models for learning.* New York: Wiley, 1955.

[8] Estes, W. K. and Burke, C. J. A theory of stimulus variability in learning. *Psychol. Rev.*, 1953, **60**, 276–286.
[9] Estes, W. K. and Burke, C. J. Application of a statistical model to simple discrimination learning in human subjects. *J. exp. Psychol.*, 1955, **50**, 81–88.
[10] Estes, W. K., Burke, C. J., Atkinson, R. C., and Frankmann, J. P. Probabilistic discrimination learning. *J. exp. Psychol.*, 1957, **54**, 233–239.
[11] Feller, W. *An introduction to probability and its applications.* New York: Wiley, 1950.
[12] Fréchet, M. *Recherches théoriques modernes sur le calcul des probabilités*, Vol. 2. Paris: Gauthier-Villars, 1938.
[13] Kemeny, J. G. and Snell, J. L. Markov processes in learning theory. *Psychometrika*, 1957, **22**, 221–230.
[14] Popper, J. and Atkinson, R. C. Discrimination learning in a verbal conditioning situation. *J. exp. Psychol.*, 1958, **56**, 21–25.
[15] Restle, F. A theory of selective learning with probable reinforcements. *Psychol. Rev.*, 1957, **64**, 182–191.
[16] Wyckoff, L. B., Jr. The role of observing responses in discrimination behavior. *Psychol. Rev.*, 1952, **59**, 437-442.

Manuscript received 1/10/58
Revised manuscript received 5/30/58

PSYCHOMETRIKA—VOL. 23, NO. 4
DECEMBER, 1958

REMARKS ON THE TEST OF SIGNIFICANCE FOR THE METHOD OF PAIRED COMPARISONS*

R. DARRELL BOCK
UNIVERSITY OF CHICAGO†

A three-component model for comparative judgment which allows for individual differences in preference is proposed. An implication of the model is that errors in the observed proportions due to sampling individuals in paired comparisons experiments are correlated. By neglecting this correlation, Mosteller's test for the method of paired comparisons tends to accept falsely the goodness of fit of the Case V solution. It is shown that bounds may be set for the correlation effect which make a valid test possible in some cases and provide useful standard errors for the estimated affective values.

As a test of the goodness of fit of the model underlying a paired comparisons solution, Thurstone [9] compared the observed proportions with those reconstructed from the solution. If discrepancies between the observed and derived proportions were small in practical terms, he considered the solution internally consistent. In 1951, Mosteller [7] suggested the use of the arcsine transformation for proportions to test whether the variance of those discrepancies is in excess of that expected from the binomial sampling variability of the observed proportions. Unfortunately, when this test is applied to data from moderate sized samples, it persistently shows that the discrepancies are *smaller* than those expected from sampling variability. That is, the fit of the (Case V) model appears to be too good rather than too poor. Mosteller's example shows this effect, as do most of the results reported by Bliss [1], and in working with preference data the present author has encountered it repeatedly.

It is shown in this paper that the anomalous behavior of Mosteller's test is the result of assuming the sampling errors independent when in general they are not. In brief, it is concluded that (i) under Case V assumptions, sampling errors for comparisons involving a common object are correlated and on certain assumptions, bounds for this correlation may be set, and (ii) with a minor alteration, Mosteller's test may be recast in the form of an analysis of variance and the effect of correlation on the variance due to departure from the Case V solution may be derived. It is then apparent that the binomial sampling variance used as the error term by Mosteller is in

*Preparation of this paper has been supported in part by the Quartermaster Food and Container Institute for the Armed Forces. Views and conclusions expressed herein are those of the author and do not necessarily reflect the views or endorsement of the Department of Defense. Comments of one of the reviewers, which substantially improved an earlier version of this paper, are gratefully acknowledged.

†Now at the University of North Carolina.

general too large and that the test must frequently fail to detect departure from internal consistency in the Case V solution.

A Three-Component Model for Comparative Judgment

Each of N individuals is required to choose the object which he prefers in each of the possible pairs formed from n objects. Individual h thus makes $n(n-1)/2$ comparisons in some order, and in the tth comparison, which consists, say, of objects i and j, his choice is assumed to be determined by the momentary affective values that the objects have for him. These values have the composition

$$(1) \qquad \begin{aligned} Y_{hit} &= \mu_i + \nu_{hi} + \epsilon_{hit} , \\ Y_{hjt} &= \mu_j + \nu_{hj} + \epsilon_{hjt} , \end{aligned}$$

where μ_i , μ_j are components fixed for specific objects and common to all individuals. These components are responsible for the concordance in preference among the N individuals.

ν_{hi} , ν_{hj} are components peculiar to specific objects and individuals, and random over the sample of individuals. These components are responsible for the consistent differences among the preferences of different individuals. Their distribution due to sampling individuals is bivariate normal $N(0, 0, \zeta_i , \zeta_j , \rho_{ij})$ for all i, j. Assume Thurstone's Case V model, as amended by Guttman [5] so that $\zeta_i = \zeta_j$ and $\rho_{ij} = \rho$ for all objects.

Finally, ϵ_{hit} and ϵ_{hjt} are error components which affect randomly the momentary judgments of each individual and result in lack of transitivity in the preferences (the circular triads of Kendall [6]). Over individuals they are assumed to be independently distributed as $N(0, \delta)$ for all objects.

It is supposed that object i is preferred to j when $Y_{hit} > Y_{hjt}$. Alternatively, define the difference

$$(2) \qquad X_{hij} = Y_{hit} - Y_{hjt} = \mu_i - \mu_j + \nu_{hi} - \nu_{hj} + \epsilon_{hit} - \epsilon_{hjt} ,$$

so that i is preferred to j when $X_{hij} > 0$. The subscript t on X_{hij} has been dropped because subscripts for the objects involved, i, j, are sufficient to identify the comparison.

From the Case V assumptions,

$$\begin{aligned} \underset{h}{E}(X_{hij}) &= \mu_i - \mu_j , \\ \underset{h}{\operatorname{var}}(X_{hij}) &= 2\zeta^2(1-\rho) + 2\delta^2 . \end{aligned}$$

Also for two comparisons sharing a common object, say i, j, and i, k,

$$\begin{aligned} \underset{h}{\operatorname{cov}}(X_{hij} , X_{hik}) &= \zeta^2(1 - \rho_{ij} - \rho_{ik} + \rho_{kj}) \\ &= \zeta^2(1-\rho), \end{aligned}$$

while for two comparisons not sharing a common object, say, i, j, and k, l,

$$\begin{aligned}\text{cov}\,(X_{hij}\,,\,X_{hkl}) &= \zeta^2(\rho_{ik} - \rho_{ij} - \rho_{kl} + \rho_{ki}) \\ &= 0.\end{aligned}$$

Since judgments of different individuals are independent, the variances and covariances of the mean differences taken over samples of N individuals are $1/N$ times the variances and covariances for the single differences. Accordingly, the correlation from one sample to another of mean differences involving a common object is

$$\rho_X = \frac{\zeta^2(1 - \rho)}{2\zeta^2(1 - \rho) + 2\delta^2}. \tag{3}$$

The ratio of variance in differences contributed by lack of transitivity within individuals to that contributed by differences in preference which are encountered in sampling individuals is defined as

$$r = \frac{\delta^2}{\zeta^2(1 - \rho)}.$$

In terms of this definition (3) may be expressed as

$$\rho_X = \frac{1}{2 + 2r}.$$

It may be observed that when individuals differ in their preferences but are perfectly transitive in their judgments, the correlations of the mean affective differences involving a common object take on their maximum value of 1/2. When the departure of the preferences of the individuals from perfect concordance is due only to intransitivity, the correlations take on their minimum value of zero.

Relationship of the Observations to the Model

The observations in the conventional paired comparisons experiment consist of the expressed preference of each individual for one of the objects in each of the $n(n - 1)/2$ pairs. For the pairs of objects i, j, and i, k, the preferences of a single individual h may be represented by the formal variates

$$t_{ij} = \begin{cases} 1 & (i \text{ preferred to } j) \\ 0 & (j \text{ preferred to } i), \end{cases}$$

$$t_{ik} = \begin{cases} 1 & (i \text{ preferred to } k) \\ 0 & (k \text{ preferred to } i). \end{cases}$$

For randomly chosen individuals, t_{ij} and t_{ik} may be considered stochastic

variates, independent from individual to individual, with probability distribution

$$P(t_{ij} = 1) = P_{ij}\ , \qquad P(t_{ij} = 0) = (1 - P_{ij}),$$
$$P(t_{ik} = 1) = P_{ik}\ , \qquad P(t_{ik} = 0) = (1 - P_{ik}),$$
$$P\{(t_{ij} = 1) \quad \text{and} \quad (t_{ik} = 1)\} = P_{ij,ik}\ .$$

For a sample of N individuals, the observed proportions of preferences of i to j and i to k are

$$p_{ij} = (\sum t_{ij})/N,$$
$$p_{ik} = (\sum t_{ik})/N,$$

where the summation is understood to be over the N individuals.

Since the individuals are selected independently, for samples of size N,

$$E(p_{ij}) = P_{ij}$$
$$E(p_{ik}) = P_{ik}$$
$$\text{var}\ (p_{ij}) = P_{ij}(1 - P_{ij})/N$$
$$\text{var}\ (p_{ik}) = P_{ik}(1 - P_{ik})/N$$
$$\text{cov}\ (p_{ij}\ , p_{ik}) = (P_{ij,ik} - P_{ij}P_{ik})/N$$

(see [3], p. 192). The observations may, in fact, be regarded as coming from a bivariate binomial process with parameters N, P_{ij} , P_{ik} , and $P_{ij,ik}$. For this distribution the above results are well known ([6], p. 133).

The Thurstone solutions for paired comparisons data assume that the expected proportions are connected with the mean affective difference $\bar{X}_{ij}$ by the normal response law

$$P_{ij} = \frac{1}{\sqrt{2\pi}} \int_{-\bar{X}_{ij}/\gamma}^{\infty} \exp\ (-y^2/2)\ dy, \qquad (-\infty < \bar{X}_{ij} < \infty)$$

where $\gamma^2 = 2\zeta^2(1 - \rho) + 2\delta^2$ and $\bar{X}_{ij} = E(X_{ij}) = \mu_i - \mu_j$. Consequently, the normal deviate corresponding to the sample proportion, p_{ij} , is taken as an estimate of $\bar{X}_{ij}/\gamma$.

For statistical purposes the use of normal deviates as estimates of the standardized mean affective differences is not convenient. The sampling variances of the deviates depend upon the population proportions, P_{ij} , and cannot be assumed constant as required for a simple least squares solution and analysis of variance. This difficulty can be avoided by departing slightly from the Thurstone solution and assuming the angular response law

$$P_{ij} = \int_0^{(\bar{X}_{ij}+c)} \sin\ (fy)\ dy, \qquad (0 \leq f(\bar{X}_{ij} + c) \leq \pi) \tag{4}$$

where c and f are location and scale constants which determine the mean and variance of the response distribution. For example, if $\tilde{X}_{ij}/\gamma$ ranges between $\pm$ $(3/4)\pi$, and $c = (3/4)\pi$ and $f = 2/3$, the mean of the response distribution is $(3/4)\pi$ and the variance is $(9/16)\pi^2 - 9/2 = 1.0516\cdots$.

Since the estimates of affective value from the paired comparison solution are unique only to a linear transformation, these constants may be incorporated into the transformation based on (4) and the angles

$$x_{ij} = 2\sin^{-1}\sqrt{p_{ij}} - \pi/2, \qquad (-\pi/2 \leq x_{ij} \leq \pi/2) \tag{5}$$

defined as estimates of the mean affective differences with arbitrary unit. The form (5) for the angular transformation is convenient because the angles center about zero and their sampling variance is nearly stable at $1/N$. Except for the arbitrary unit, this transformation closely parallels the normal in the range from $P = .05$ to $.95$ [4]. If the paired comparisons solution is confined to proportions in or near this interval, the finite range of the x_{ij} will present no theoretical difficulties. Furthermore, the results of this paper will apply to a good degree of approximation to the Case V solution based on the normal response law.

For two comparisons sharing a common object, moments of the asymptotic distribution of the angles from samples of size N may be obtained by expanding the right-hand member of (5) in a Taylor's series about P_{ij}. Neglecting terms in which $(p_{ij} - P_{ij})$ appears in degree higher than the first, and taking term by term expectations of the series, their squares, and the product of the two series,

$$\begin{aligned} E(x_{ij}) &= 2\sin^{-1}\sqrt{P_{ij}} - \pi/2 \\ E(x_{ik}) &= 2\sin^{-1}\sqrt{P_{ik}} - \pi/2 \\ \text{var}\,(x_{ij}) &\cong 1/N \\ \text{var}\,(x_{ik}) &\cong 1/N \\ \text{cov}\,(x_{ij}, x_{ik}) &\cong \frac{(P_{ij,ik} - P_{ij}P_{ik})}{N\sqrt{P_{ij}(1-P_{ij})P_{ik}(1-P_{ik})}}. \end{aligned} \tag{6}$$

Whence, the sampling correlation of the observed angles is

$$\rho_{x_{ij},x_{ik}} = \frac{(P_{ij,ik} - P_{ij}P_{ik})}{\sqrt{P_{ij}(1-P_{ij})P_{ik}(1-P_{ik})}}.$$

According to the three-component model, $P_{ij,ik}$ will not in general equal $P_{ij}P_{ik}$ since the comparisons share the common object i. As a result, the sampling correlation of the angles within rows and columns of the paired comparisons table will not vanish as required for a conventional analysis of variance. In order to obtain bounds for the magnitude of this correlation ef-

fect on Case V assumptions, it is of interest to study $\rho_{x_{ij},ik}$ as a function of the parameters of the three-component model. Let $\gamma = 1$, so that the distribution of X_{ij} , X_{ik} is

$$N(\bar{X}_{ij} , \bar{X}_{ik} , 1, 1, \rho_X).$$

Then the marginal proportions P_{ij} and P_{ik} can be obtained by entering the table of the normal distribution with $\bar{X}_{ij}$ and $\bar{X}_{ik}$ respectively. The joint proportion $P_{ij,ik}$ can be obtained from Pearson's table of the bivariate normal distribution, entering with $\bar{X}_{ij}$, $\bar{X}_{ik}$, and ρ_X [8]. Values of $\rho_{x_{ij},ik}$ have been calculated for selected values of these parameters and are shown graphically in Figures 1 and 2.

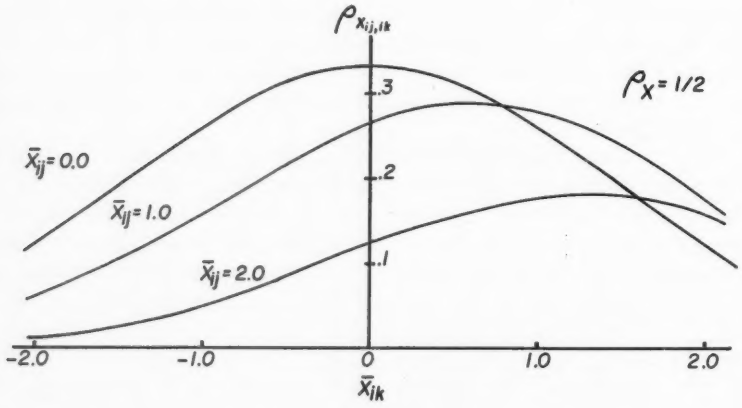

FIGURE 1

The correlation of the observed angles as a function of the mean affective differences. (Negative values of $\bar{X}_{ij}$ yield a similar figure reversed from right to left.)

It is clear from Figure 1 that the sampling covariance of the observed angles where a common object is involved, unlike the variance, is not independent of the values assumed for the mean affective differences. If the correlation ρ_X for the three-component model is assumed constant at 1/2, $\rho_{x_{ij},ik}$ reaches a maximum of 1/3 when $\bar{X}_{ij} = \bar{X}_{ik} = 0$, that is, when a 50 per cent split is expected in the preferences for the two comparisons. As $\bar{X}_{ij}$ and $\bar{X}_{ik}$ depart from one another, so that the preferences for the common object approach 100 per cent in one of the comparisons, the correlation for the observed angles falls to near zero. In Figure 2, $\rho_{x_{ij},ik}$ is seen to decrease almost linearly with ρ_X from 1/3 to 0 when $\bar{X}_{ij}$ and $\bar{X}_{ik}$ are zero. For other values of the means the change of $\rho_{x_{ij},ik}$ with ρ_X is restricted and more curvilinear.

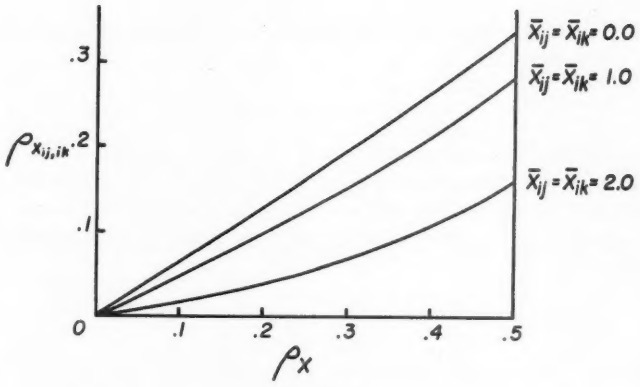

FIGURE 2

The correlation of the observed angles as a function of the correlation of the mean affective differences.

In order to obtain a workable solution for the three-component model it is necessary to assume that $\rho_{x_{ij,ik}}$ is constant and equal to, say ρ_x, in all comparisons. In general, this is not a good assumption, even if ρ_X can be assumed constant, because when the mean affective values of the objects differ greatly there will be many comparisons for which $\rho_{x_{ij,ik}}$ is reduced by extreme values for P_{ij} and P_{ik}. On the other hand, in precisely those cases where the full efficiency of a least squares solution is required and statistical significance is in question, i.e., when the objects differ little in affective value, the proportions of all comparisons will be near enough to 50 per cent for the assumption of constant $\rho_{x_{ij,ik}}$ to be reasonable. The simple solution and sampling criteria which result from this assumption are therefore of practical interest.

An Analysis of Variance for the Case V Solution

Let the observed angles (5) for a paired comparisons experiment be set out in the form of (7).

(7)

Objects	1	2	⋯	n	Sums
1	0	x_{12}	⋯	x_{1n}	$x_{1\cdot}$
2	x_{21}	0	⋯	x_{2n}	$x_{2\cdot}$
⋮	⋮	⋮		⋮	⋮
n	x_{n1}	x_{n2}	⋯	0	$x_{n\cdot}$
Sums	$x_{\cdot 1}$	$x_{\cdot 2}$	⋯	$x_{\cdot n}$	0

$x_{ij} = -x_{ji}$

$x_{i\cdot} = -x_{\cdot i}$

Since the grand mean for (7) is exactly zero, consider the elements of (7) to have the composition

$$x_{ij} = \alpha_i + \beta_j + \epsilon_{ij} \,. \tag{8}$$

That is, the row and column effects of (7) are considered additive. Then, letting $\beta_i = -\alpha_i$, the observational equation (8) is of the simple subtractive form specified by the model (2). Assuming the variance and correlation for the error term are constant and taking account of the skew symmetry of (7), for distinct h, i, j, k

$$\begin{aligned} E(\epsilon_{ij}) &= 0, \\ E(\epsilon_{ij}^2) &= \sigma^2, \\ E(\epsilon_{ij}\epsilon_{ji}) &= -\sigma^2, \qquad (\sigma^2 = 1/N) \\ E(\epsilon_{ij}\epsilon_{ik}) &= \rho_x\sigma^2, \\ E(\epsilon_{ij}\epsilon_{ki}) &= -\rho_x\sigma^2, \\ E(\epsilon_{hi}\epsilon_{jk}) &= 0. \end{aligned} \tag{9}$$

Writing (8) in terms of sample estimates

$$x_{ij} = a_i + b_j + e_{ij}$$

and minimizing the correlated error demonstrates that the correlation terms drop out, because ρ_x is constant, and the normal equation (10) results.

$$-\sum_j x_{ij} + \sum_j b_j + na_i = 0. \tag{10}$$

Because of the skew symmetry of (7), $b_i = -a_i$; then assuming $\sum_i b_i = 0$, as in a conventional analysis of variance,

$$a_i = \sum_j x_{ij}/n \,.$$

Because of the equivalence of (2) and (8), these a_i may be regarded as estimates of the affective values of the objects, up to a linear transformation. This means that Thurstone's Case V solution in terms of arcsines is a least squares solution even under the assumption of constantly correlated error within rows and columns of the paired comparisons table.

For the analysis of variance associated with this solution the expectations of the row, column, and residual sums of squares must be calculated assuming the covariance structure for error given by (9). The sum of squares between rows of (7) taken about the mean for the table (which is exactly zero) is

$$\text{SS}A = \frac{1}{n}\sum_i x_{i\cdot}^2.$$

$$= \frac{1}{n}\sum_i (n\alpha_i + \sum_j \epsilon_{ij})^2$$

$$= \frac{1}{n}[\sum_i n^2\alpha_i^2 + 2n\sum_i \alpha_i(\sum_j \epsilon_{ij}) + \sum_i (\sum_j \epsilon_{ij})^2].$$

Since the diagonal entries in (7) are without error, the expected sum of squares is

$$E(\text{SS}A) = n\sum_i \alpha_i^2 + (n-1)\sigma^2 + (n-1)(n-2)\rho_z\sigma^2.$$

Similarly, the expected sum of squares between columns, which is numerically equal to that between rows, is

$$E(\text{SS}B) = n\sum_j \beta_j^2 + (n-1)\sigma^2 + (n-1)(n-2)\rho_z\sigma^2.$$

The expected sum of squares for the residual may be obtained by subtraction or derived as follows.

$$\text{SS}R = \sum_i \sum_j \left(x_{ij} - \frac{x_{i\cdot}}{n} - \frac{x_{\cdot j}}{n}\right)^2.$$

Since $\sum_i x_{i\cdot} = 0$ and $\sum_j x_{\cdot j} = 0$,

$$\text{SS}R = \sum_i \sum_j x_{ij}^2 - \sum_i x_{i\cdot}^2/n - \sum_j x_{\cdot j}^2/n.$$

Then

$$\text{SS}R = \sum_i \sum_j (\alpha_i + \beta_j + \epsilon_{ij})^2$$

$$- \frac{1}{n}\sum_i (n\alpha_i + \sum_j \epsilon_{ij})^2 - \frac{1}{n}\sum (n\beta_j + \sum_i \epsilon_{ij})^2$$

$$= \sum_i \sum_j \epsilon_{ij}^2 - \frac{1}{n}\sum_i (\sum_j \epsilon_{ij})^2 - \frac{1}{n}\sum_j (\sum_i \epsilon_{ij})^2;$$

$$E(\text{SS}R) = (n-1)(n-2)\sigma^2 - 2(n-1)(n-2)\rho_z\sigma^2.$$

These results are collected in Table 1.

It may be noted that

$$\text{var}\,(a_i) = \frac{n-1}{n^2}[1 + (n-2)\rho_z]\sigma^2;$$

$$\text{cov}\,(a_i, a_j) = -\frac{1}{n^2}[1 + (n-2)\rho_z]\sigma^2; \tag{11}$$

$$\text{var}\,(a_i, -a_j) = \frac{2}{n}[1 + (n-2)\rho_z]\sigma^2.$$

TABLE 1

Analysis of Variance for the Case V Solution *

(Assuming a three-component model and the arcsine transformation of the observed proportions)

Source of variation	d.f.	Sums of squares	Expected sums of squares
Between objects	$n - 1$	$SSA = \frac{1}{n}\Sigma x_{i.}^2$	$E(SSA) = (n-1)\left[1+(n-2)\rho_x\right]\sigma^2 + n\Sigma\alpha_i^2$
Residual	$(n-1)(n-2)/2$	$SSR = SST - SSA$	$E(SSR) = \frac{1}{2}(n-1)(n-2)(1-2\rho_x)\sigma^2$
Total	$n(n-1)/2$	$SST = \Sigma\Sigma_{i<j} x_{ij}^2$	$\left[\sigma^2 = 1/N\right]$

* Based on values in the upper half of the paired comparisons table only.

Discussion

Except for the approximation of the normal response law by the angular, Mosteller's test for the goodness of fit of the Case V solution is equivalent to a χ^2 test, on $(n - 1)(n - 2)/2$ degrees of freedom, calculated by dividing the residual sum of squares in Table 1 by σ^2. If ρ_z is greater than zero, however, the expectation of this χ^2 is not $(n - 1)(n - 2)/2$ but is diminished by a factor of $(1 - 2\rho_z)$. According to the results established previously, ρ_z approaches 1/3 if the preferences of the individuals are completely transitive, the P_{ij} approach 1/2, and N is large enough to justify the approximations in (6). Under these conditions the χ^2 for the test of goodness of fit would be signficantly *small* if the data conform to the model. The three-component model is, therefore, able to account for the aberrant behavior of Mosteller's test. When the individuals are not transitive in their preferences or there are many extreme proportions of preference, the reduction of the residual χ^2 would be less apparent.

Similarly, a χ^2 on $(n - 1)$ degrees of freedom for testing signficance of differences in the affective values of the objects, if calculated from SSA/σ^2, is augmented in expectation by a factor of $[1 + (n - 2)\rho_z]$. Thus, when correlated error is present and its effect ignored, the investigator will be more confident than is warranted about the significance of differences among and the stability of the estimated affective values.

It has been shown by Walsh [10] that correlation of the observations can be considered to alter the effective variance of sampling error. Its effect can be nullified, however, by incorporating the altered variance in the sampling statistics derived from the error distribution. Thus, the correct χ^2 statistics

for Table 1 are,

$$\chi_R^2 = \frac{SSR}{(1 - 2\rho_z)\sigma^2}, \qquad \text{d.f.} = (n - 1)(n - 2)/2, \tag{12}$$

$$\chi_A^2 = \frac{SSA}{[1 + (n - 2)\rho_z]\sigma^2}, \qquad \text{d.f.} = (n - 1), \tag{13}$$

for the residual and between-object variance respectively.

Since there appears to be no very satisfactory way of estimating ρ_z or even assuring its constancy, the corrected χ^2's and the estimated variances (11) cannot provide exact tests of significance or standard errors. However, bounding conditions which should be useful in some applications may be established. For example if ρ_z is assumed to take on its maximum value, 1/3, a residual χ^2 with the correlation effect cancelled may be computed from (12). If this χ^2 is not significant, we can be confident that there is no evidence of departure from the model. Conversely, if ρ_z is assumed equal to zero and the χ^2 is significant, there is evidence of departure. For intermediate cases no conclusion can be drawn.

Similarly, if χ^2 for between-object variation is significant when computed from (13) setting $\rho_z = 1/3$, there is evidence of significant discrimination. Conversely, if this χ^2 is not significant when $\rho_z = 0$ is used, there is no evidence of significant discrimination among the objects. Intermediate results remain uncertain.

For standard errors of the estimated affective values, it should suffice in many cases to obtain upper bounds by assuming the condition most unfavorable to discrimination among the objects, namely, $\rho_z = 1/3$. Substituting this value in (11) and taking square roots gives *maximal* standard errors for the estimated values and their differences.

Finally, it should be pointed out that for paired comparisons obtained from the repeated judgments of the same individual and with the identity of the objects concealed, as, for example, in organoleptic testing, ρ_z could probably be assumed to vanish. This assumes that there is no replication effect specific to particular objects. In this case, the expected sums of squares in Table 1 would reduce to their ordinary form and exact tests of the residual and between object variance would be possible. Thus, paired comparisons methods in organoleptic testing, such as Bradley's [2], are valid when confined to repeated judgments of the same individual, but would encounter the same difficulties as Mosteller's test if applied to group data.

REFERENCES

[1] Bliss, C. I., Greenwood, M. L., and White, E. S. A rankit analysis of paired comparisons for measuring the effects of sprays on flavor. *Biometrics*, 1956,12, 381–403.

[2] Bradley, R. A. and Terry, M. E. The rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika*, 1952, 39, 324–345.

[3] Cramer, H. *Mathematical methods of statistics.* Princeton: Princeton Univ. Press, 1951.
[4] Finney, D. J. *Statistical method in biological assay.* London: Griffin, 1952.
[5] Guttman, L. An approach for quantifying paired comparisons and rank order. *Ann. math. Statist.*, 1946, **17**, 144–163.
[6] Kendall, M. G. *The advanced theory of statistics.* Vol. I. London: Griffin, 1948.
[7] Mosteller, F. Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviation and equal correlations are assumed. *Psychometrika*, 1951, **16**, 207–218.
[8] Pearson, K. *Tables for statisticians and biometricians.* Part II. London: Univ. London, 1931.
[9] Thurstone, L. L. The method of paired comparisons for social values. *J. abnorm. soc. Psychol.*, 1927, **21**, 384–400.
[10] Walsh, J. E. Concerning the effect of intraclass correlation on certain significance tests. *Ann. math. Statist.*, 1947, **18**, 88–96.

Manuscript received 5/22/57
Revised manuscript received 3/11/58

PSYCHOMETRIKA—VOL. 23, NO. 4
DECEMBER, 1958

A COMPARISON OF THE PRECISION OF THREE EXPERIMENTAL DESIGNS EMPLOYING A CONCOMITANT VARIABLE

LEONARD S. FELDT
STATE UNIVERSITY OF IOWA

Three techniques are commonly employed to capitalize on a concomitant variate and improve the precision of treatment comparisons: (1) stratification of the experimental samples and use of a factorial design, (2) analysis of covariance, and (3) analysis of variance of difference scores. The purpose of this paper is to compare the effectiveness of these alternatives in improving experimental precision, to identify the most precise design and the conditions under which its advantage holds, and to derive, in the case of the factorial approach, recommendations as to the optimal numbers of levels.

In order to improve the precision of what would otherwise be a completely randomized analysis of variance design, educational and psychological experimenters often consider designs which involve the use of a concomitant or control variable. Such a variable is usually defined by a characteristic of the subjects which more or less predetermines the general level of their criterion measure and is highly correlated with it. For example, test intelligence is frequently used as a control variable in educational experiments on teaching methods, since a high correlation is known to exist between intelligence test scores and the scores on the achievement tests typically used as criterion measures in such experiments. Three techniques are commonly used to improve the precision of the experimental design: two-factor analysis of variance with two or more observations per cell, analysis of covariance, and analysis of variance of difference scores. The purpose of this paper is to compare the effectiveness of these alternatives and, in the case of the factorial approach, to derive recommendations as to the optimal numbers of levels under various conditions.

In the factorial or treatments-by-levels approach, as it has been called by Lindquist [17], levels or intervals are defined along the scale of values of the control variable, and subjects within any level are assigned to treatments at random. In almost all cases in which the treatments are superimposed by the experimenter the subjects are assigned to the several treatments in the same proportions for the various levels in order to simplify the analysis. The experimenter has, as a rule, little intrinsic interest in the main effects of the control variable. The assumption usually can be made that the subpopulations corresponding to the various levels within any treatment differ in their mean criterion measure. It is also often expected that there will be no interaction between the levels of the control variable and the treatments variable;

however, the absence of interaction cannot be assumed. On many occasions a test for the presence of interaction constitutes a secondary purpose of the experiment. Because both the levels and treatments variables are assumed to be fixed, nonrandom effects, the design is generally set up with at least two observations per cell, to make available an estimate of error variance that does not include the levels and interaction effects. Thus the purpose of the levels factor in the two-factor analysis is primarily to stratify the samples assigned to the other treatment categories. To the extent of the relationship between the control and criterion measures in the experimental population, such stratification results in control of an important source of error variance and hence improves the precision of the experiment.

Analysis of covariance provides a second alternative by which a potential source of error variance may be controlled. In place of the stratification of the experimental samples to reduce random differences between treatment groups, regression equations are used to adjust criterion measure differences among subjects, to the extent that these differences are associated with control measure differences. Before the technique may be applied some assumption must be made as to the nature of the relationship between the control and criterion variables and the homogeneity of this relationship from treatment to treatment. If these assumptions are fulfilled, any experiment designed to permit valid analysis as a completely randomized design will have a valid analysis via covariance techniques.

The essential feature of the method of differences is the definition of the criterion. In the factorial and covariance approaches the control variable score X and a criterion variable score Y are not combined; under the difference method the criterion measure is defined as $(X - Y)$ or $(Y - X)$. These data are then analyzed as a completely randomized design. This technique is probably most frequently employed in cases where X and Y may be considered parallel forms of a test. For example, in educational experimentation, the X or control variable is often defined as a pretest administered before the initiation of the experimental treatments, and the Y variable is defined by a final test administered after completion of the treatments. Rarely would an experimenter be inclined to use such a difference score unless it made intuitive sense as a measure of change in performance or gain in skill.

These three alternatives will be considered as they apply to the one-factor, completely randomized design involving t treatments, with an independent random sample assigned to each treatment. The number of experimental subjects, N, will be divided equally among the t treatments; that is, $N = tn$. Assume that these samples are drawn from t normally distributed populations with equal variances. A continuous control variable X is available which is linearly related to Y, the criterion measure. The population mean and variance of X and the population value of the correlation coefficient, ρ,

between X and Y, are assumed equal for all t treatment populations. Finally, assume that homoscedasticity in Y obtains around the regression line of Y on X. The experimental situation here assumed is, therefore, one in which the conditions of the completely randomized design and covariance are fully satisfied. It is also typical of the situations in which the two-factor and difference methods are employed.

Related Research

The advantages of analysis of covariance over analysis of variance, when all assumptions of both models are satisfied and a highly correlated control variable is available, have been emphasized by writers in many experimental fields. The statistical literature includes many empirical demonstrations of the reduction in error variance accomplished through statistical control of a pertinent variable. Considerably less attention has been paid to the *comparative* efficiencies of other techniques which also capitalize on concomitant information.

Fisher [11, 12] develops and illustrates the three techniques under consideration here, but does not compare them rigorously with respect to precision. Other popular texts [1, 5, 8, 16, 21] follow much the same pattern. Discussion of the efficiency of covariance designs is generally limited to a comparison with the corresponding analysis of variance.

Federer [8] raises the problem of covariance versus stratification and suggests various experimental situations in which both techniques might be useful. While he makes no systematic study of their comparative precision, Federer clearly favors stratification. He suggests the following rule to experimenters: "If the experimental variation cannot be controlled by stratification, then measure related variates and use covariance" ([8], pp. 483–484). This view appears to be shared by most of the workers in the field. The opinion is based on a number of considerations. (i) It is generally accepted that the differences in the precision of the various designs are relatively small, even for moderate numbers of degrees of freedom. (ii) The greater number of assumptions required for a valid analysis of covariance renders the technique less generally applicable. (iii) The experimenter is often forced to make the crucial choice of a regression model on rather tenuous bases. (iv) The failure of the data to meet the assumptions of the model is thought to be more serious in covariance than in regular analysis of variance, especially with respect to failure to satisfy regression assumptions. (v) The danger exists that the effects which are eliminated may actually be relevant to the objectives of the treatments.

Outhwaite and Rutherford [20] give empirical evidence which suggests that when the number of replicates per treatment is approximately equal to the number of treatments, a modified Latin square design is more efficient than a covariance analysis which takes into account all possible higher order

regressions. Essentially, they conclude that in their case stratification on two variables results in a more precise experiment than stratification on one and statistical control on the other.

Lucas [18] has considered covariance designs in which treatment groups were balanced with respect to the mean value of the control variate. Essential to his discussion is the following expression for the expected value of the adjusted mean square for treatments in a covariance analysis.

$$E[\text{ms}_t^*] = \sigma^2 + n\left[1 - \frac{T_{xx}}{(t-1)S_{xx}}\right]\left[\frac{\sum(\mu_i - \mu)^2}{t-1}\right].$$

In this expression, which assumes fixed treatment effects,

ms_t^* = adjusted mean square for treatments,
n = number of replicates per treatment,
t = number of treatments,
T_{xx} = sum of squares between treatments on X,
S_{xx} = total sum of squares on X,
$\sum(\mu_i - \mu)^2$ = sum of squared deviations of the treatment population means from the mean for all treatment populations.

It is assumed that the treatment groups are random samples from a common population on X. Lucus suggests that the term $T_{xx}/[(t-1)S_{xx}]$ may be used as a measure of the loss in sensitivity due to failure to achieve perfect balancing on X. He then indicates that for experiments involving small numbers of degrees of freedom, a slight gain in efficiency may be obtained by a sampling procedure which achieves balancing. These conclusions were reiterated by Greenberg [15].

Gourlay [13] compared the techniques of stratification, covariance, and the analysis of variance of differences and concluded that covariance always results in the most precise experiment. However, Gourlay failed to consider the sampling error involved in the estimate of β and disregarded the effect of differences in degrees of freedom for error. Thus, in his discussion the error variance of a single adjusted mean under a covariance analysis was $\sigma_Y^2(1 - \rho^2)/n$, where σ_Y^2 is the variance of the criterion measure within any treatment population, ρ the correlation between criterion and control measures, and n the number of replicates per treatment. Such a value applies only for the case in which the treatment sample mean on the control measure equals the general mean on that variable, a condition which does not generally hold in strict random sampling. All comparisons with this value naturally indicated a difference in precision which favored covariance. Gourlay dismissed the factorial or levels design without deriving any formal expression for its precision, since the error variance of a treatment mean clearly approached $\sigma_Y^2(1 - \rho^2)/n$ only as a limit.

Cox [6] made an extensive study of various techniques for employing

concomitant information in an experimental design. He employed two measures of experimental imprecision. The first, which he called the true imprecision of the experiment, was based on the population value of the average error variance for the difference between two treatment means, adjusted by covariance where appropriate. The second, which he called the apparent imprecision of the experiment, was defined as the product of the true imprecision times an adjustment factor based on the degrees of freedom for error. This adjustment was originally proposed by Fisher [12]. It makes possible a more meaningful comparison of the relative efficiency of two experiments which utilize the same total number of subjects but which give rise to unequal numbers of degrees of freedom for error. Symbolically, these measures are defined as follows.

$$I_t = \text{average var } (M_i - M_k)/(2\sigma_0^2/n);$$

$$I_a = I_t\left(\frac{f+3}{f+1}\right).$$

In these expressions f is the degrees of freedom for error, σ_0^2 is the variance of Y for fixed X; $2\sigma_0^2/n$ is the theoretical minimum for the variance of the difference between two treatment means. Thus $I_a > I_t \geq 1.00$.

Cox evaluated I_t and I_a for the covariance and factorial designs for a number of combinations of N, ρ, and t, the number of treatments. He concluded that stratification is more advantageous for $\rho < .6$ and that covariance becomes appreciably better than the block design only when ρ is as large as .8 or more. He noted further that the block design is reasonably efficient for any form of smooth regression, not just for linear regression. However, if the distribution of X is leptokurtic, the efficiency of the block design is lowered due to the end blocks having units with widely discrepant values of X.

The block design in Cox's discussion is formulated by ranking subjects on X, subdividing the ranked subjects in groups of t each, and assigning one subject per block at random to each treatment. The interaction of blocks with treatments is then used as the error term. Such a design can rarely be used in psychological or educational experimentation, since an a priori assumption of no intrinsic interaction can rarely be made. In these fields the blocks are not generally selected randomly, and the main effects and interaction are regarded as fixed effects. The design is almost always set up with two or more subjects per cell to make available an error estimate based on within-cells variation. Such an estimate makes possible a test of the significance of the interaction, an effect which often has considerable experimental importance. The within-cells error estimate does not eliminate the problems of inference which arise when interaction is present, of course. As indicated in the final section of this paper, interaction is equivalent to heterogeneity of regression in analysis of covariance. It is pertinent, however, to examine

the accuracy of Cox's recommendations when applied to this type of block design and to extend them to experiments based on such numbers of cases as are commonly employed in educational and psychological research studies.

Index of Experimental Precision

Probably the most satisfactory index for the comparison of the precision of two experimental designs is one based on the average variance of the difference between all pairs of treatment means, adjusted by covariance where appropriate. The comparison of designs might be made in terms of the population value of these error variances. However, such a comparison fails to take into account variation from one design to another in the number of degrees of freedom available for error. Such differences reflect variation in the precision of the estimate of the error variance itself, and they become of some importance when the degrees of freedom available for error are quite small. To permit an evaluation which makes due allowance for information lost in the estimation of the error variance, the adjustment proposed by Fisher and noted in the previous section will be employed.

The minimum value for the variance of the difference between the means of treatments j and k is

$$\min \text{ var } (M_j - M_k) = 2\sigma_Y^2(1 - \rho^2)/n.$$

In this expression σ_Y^2 is the variance of the criterion measure in any treatment population. This result follows immediately from the assumption of homoscedasticity and the well-known expression for the variance of Y within any array for a given value of X.

Following Cox [6], the *true* imprecision I_t of a given experimental design is defined as the ratio of the population value of the average variance of $(M_j - M_k)$ for that design to min var $(M_j - M_k)$. The *apparent* imprecision I_a is defined as the product of the true imprecision times the adjustment factor proposed by Fisher. That is, for a design designated 1, the true imprecision is defined

$${}_1I_t = \frac{\text{ave var}_1 (M_j - M_k)}{\min \text{ var } (M_j - M_k)}.$$

The apparent imprecision of this design is defined

$${}_1I_a = {}_1I_t\left(\frac{f+3}{f+1}\right).$$

Thus for any pair of designs based on constant N, comparison of the respective values of I_a will indicate that design which will yield the most precise evaluation of the treatment effects.

Comparison of the Precision of the Three Designs

Two-factor Analysis of Variance (Design 1)

Under this setup several intervals, not necessarily of equal length, are defined along the scale of values of the control variable. All treatments are assigned subjects from the various levels in equal numbers. The general case in which h levels of the control variable are employed will be considered. The limits of the subpopulation corresponding to the lowest level are $-\infty$ and X_1. The general level i has the limits X_{i-1} and X_i. The highest level subpopulation has the lower limit X_{h-1} and the upper limit of $+\infty$.

In the two-factor design generated by the introduction of levels on X, all assumptions of the analysis of variance pertain to the distributions of Y in the subpopulations $1, \cdots, i, \cdots, h$ within each treatment. The error variance of this design represents an estimate of the variance of Y for these subpopulations, which, if the assumptions of the mathematical model are to be satisfied, must be equal. An expression for the variance of Y in the subpopulation at level i will be derived; in general this variance is not exactly equal to that at all other levels. That is, under the assumed experimental situation a small degree of heterogeneity of variance can be expected in the treatments-by-levels design. Expressions will then be derived for the average within-cell variances, I_t and I_a, and the latter will be evaluated for a variety of experimental conditions.

From the assumption of homogeneous linear regression, the slope of the regression line of Y on X within any level of the population is equal to the slope of the regression line for the entire bivariate surface. Using lower case letters to represent values specific to level i and upper case letters to represent the entire population, this relationship may be written as

$$\rho_{XY}\frac{\sigma_Y}{\sigma_X} = \rho_{xy}\frac{\sigma_y}{\sigma_x}. \tag{1}$$

From the assumption of homoscedasticity a second relationship involving the variance of Y within any X array is

$$\sigma_y^2(1 - \rho_{xy}^2) = \sigma_Y^2(1 - \rho_{XY}^2). \tag{2}$$

Solving (1) for ρ_{xy}, substituting this in (2), and solving for σ_y^2, the following expression for criterion variance within level i of the treatment population is derived.

$$\sigma_{yi}^2 = \sigma_Y^2\left[1 - \rho^2\left(1 - \frac{\sigma_{xi}^2}{\sigma_X^2}\right)\right]. \tag{3}$$

From (3) the variance of the criterion measure for any level subpopulation is seen to be a function of the variance of the control measure for this population. If the levels are defined to include subpopulations differing in

the variance of the control measure, they will differ in criterion measure variance also. Thus if a linear relationship exists between X and Y, the assumption of homogeneous cell variance will not, in general, be exactly satisfied for the two-factor or treatments-by-levels design.

The degree of heterogeneity which obtains from level to level within any treatment may be demonstrated for two common experimental situations in which the control variable is normally distributed: (i) the design in which the levels include *equal proportions* of the population, and (ii) the design in which the levels are defined by *equal intervals* along the scale of values of the control variable. The variance of a segment of the normal distribution may be evaluated through integration by parts. For the unit normal distribution, a general formula for the variance of segment i is

$$\sigma_{x_i}^2 = 1 + \frac{x_{i-1}z_{i-1} - x_i z_i}{A} - \left(\frac{z_{i-1} - z_i}{A}\right)^2,$$

where A is the area included in the segment, x_{i-1} and x_i the segment limits, and z_{i-1} and z_i the ordinates at these limits. This formula was used to solve for $\sigma_{x_i}^2$ for 2, 4, $\cdots$, 10 levels under both the equal proportion and equal interval definition of levels, and the resulting values substituted in (3). In the case of levels defined by equal intervals, the range from -3σ to $+3\sigma$ was used to establish interval limits; the lowest and highest intervals were then extended to $-\infty$ and $+\infty$, respectively. In Tables 1 and 2 these variances are presented for each level, assuming $\sigma_y^2 = 1$. Levels are numbered from lowest (number 1) to highest.

It may be seen from (3) and the values in Tables 1 and 2 that the degree of heterogeneity depends on the value of ρ and the manner in which the levels are defined. Levels defined by equal intervals give rise to variances slightly less heterogeneous than those arising from levels defined by equal proportions. In both instances the degree of heterogeneity is quite small, even for values of ρ as large as .8. For example, if $\rho = .8$, and six levels are used, the ratio of the smallest variance to the largest is 1:1.32 when the equal proportion method of constituting levels is used, 1:1.05 when the equal interval method is used. In view of the literature on the effects of small degrees of heterogeneity of variance on the F test [3, 4, 7, 14, 19] this degree of heterogeneity will probably not seriously invalidate the test.

The average variance over all levels has been computed for the various situations covered in Tables 1 and 2; these values are presented in the last line of each table. Represent this average as

$$\bar{\sigma}_y^2 = \sigma_y^2\left[1 - \rho^2\left(1 - \frac{\bar{\sigma}_x^2}{\sigma_X^2}\right)\right]. \tag{4}$$

Then write

$$\text{var}\,(M_i - M_k) = 2\bar{\sigma}_y^2/n,$$

TABLE 1

The Variance for Levels Which Include Equal Proportions of a Normal Population ($\sigma_y^2 = 1.00$)

Level	Number of levels					
number	2	4	6	8	10	∞
1	$1-.637\rho^2$	$1-.758\rho^2$	$1-.798\rho^2$	$1-.818\rho^2$	$1-.834\rho^2$	$1-1.00\rho^2$
2	$1-.637\rho^2$	$1-.963\rho^2$	$1-.976\rho^2$	$1-.981\rho^2$	$1-.984\rho^2$	$1-1.00\rho^2$
3		$1-.963\rho^2$	$1-.985\rho^2$	$1-.990\rho^2$	$1-.992\rho^2$	$1-1.00\rho^2$
4		$1-.758\rho^2$	$1-.985\rho^2$	$1-.992\rho^2$	$1-.993\rho^2$	$1-1.00\rho^2$
5			$1-.976\rho^2$	$1-.992\rho^2$	$1-.995\rho^2$	$1-1.00\rho^2$
6			$1-.798\rho^2$	$1-.990\rho^2$	$1-.995\rho^2$	$1-1.00\rho^2$
7				$1-.981\rho^2$	$1-.993\rho^2$	$1-1.00\rho^2$
8				$1-.818\rho^2$	$1-.992\rho^2$	$1-1.00\rho^2$
9					$1-.984\rho^2$	$1-1.00\rho^2$
10					$1-.834\rho^2$	$1-1.00\rho^2$
Mean	$1-.637\rho^2$	$1-.861\rho^2$	$1-.920\rho^2$	$1-.945\rho^2$	$1-.959\rho^2$	$1-1.00\rho^2$

TABLE 2

The Variance for Levels Defined by Equal Intervals Between $\pm 3\sigma$ of a Normal Population ($\sigma_y^2 = 1.00$)

Level	Number of levels					
number	2	4	6	8	10	∞
1	$1-.637\rho^2$	$1-.850\rho^2$	$1-.886\rho^2$	$1-.903\rho^2$	$1-.903\rho^2$	$1-1.00\rho^2$
2	$1-.637\rho^2$	$10.835\rho^2$	$1-.927\rho^2$	$1-.958\rho^2$	$1-.973\rho^2$	$1-1.00\rho^2$
3		$1-.835\rho^2$	$1-.920\rho^2$	$1-.956\rho^2$	$1-.972\rho^2$	$1-1.00\rho^2$
4		$1-.850\rho^2$	$1-.920\rho^2$	$1-.954\rho^2$	$1-.971\rho^2$	$1-1.00\rho^2$
5			$1-.927\rho^2$	$1-.954\rho^2$	$1-.970\rho^2$	$1-1.00\rho^2$
6			$1-.886\rho^2$	$1-.956\rho^2$	$1-.970\rho^2$	$1-1.00\rho^2$
7				$1-.958\rho^2$	$1-.971\rho^2$	$1-1.00\rho^2$
8				$1-.903\rho^2$	$1-.972\rho^2$	$1-1.00\rho^2$
9					$1-.973\rho^2$	$1-1.00\rho^2$
10					$1-.903\rho^2$	$1-1.00\rho^2$
Weighted Mean	$1-.637\rho^2$	$1-.837\rho^2$	$1-.921\rho^2$	$1-.954\rho^2$	$1-.970\rho^2$	$1-1.00\rho^2$

$$(5) \qquad {}_1I_t = \frac{1 - \rho^2[1 - (\bar{\sigma}_x^2/\sigma_X^2)]}{1 - \rho^2},$$

and

$$(6) \qquad {}_1I_a = {}_1I_t\left(\frac{N - th + 3}{N - th + 1}\right),$$

where h equals the number of levels.

From (5) and the values of Tables 1 and 2 it is clear that ${}_1I_t$, and ultimately ${}_1I_a$, depends upon the number of levels of X the experimenter employs. As the number of levels increases, the variance of the control variable within levels decreases, and ${}_1I_t$ approaches 1.00. It is of some interest to note in these tables the relative rapidity with which the numerator of (5) approaches $1 - \rho^2$.

It should not be inferred from (5) that the maximum experimental precision is achieved when the maximum number of levels is used. It is true that, other things being equal, smaller values of ${}_1I_t$ indicate a smaller population variance for $M_i - M_k$. However, in a two-factor experiment, like that considered here, additional degrees of freedom are lost from error with every addition to the number of levels. This reduction in degrees of freedom represents a loss in experimental precision, a loss that may be either less than, equal to, or greater than the gain associated with the reduction in the population variance of $M_i - M_k$. For example, a loss of four degrees of freedom from 120 to 116 represents only a negligible loss of power and may be more than justified by the decrease in the error variance derived from one or two additional levels. On the other hand, a loss of four degrees of freedom from 20 to 16 may result in a loss in power that exceeds the gain accruing from the reduction in error variance. Thus it can not be unconditionally concluded that the greater the number of levels, the greater the precision of the treatments-by-levels design. The optimal number of levels is contingent upon the total number of degrees of freedom available; it is for this purpose Fisher's adjustment for I_t is introduced.

Since reductions in the population error variance decrease monotonically with increasing numbers of levels, and since the effect of the adjustment factor $(f + 3)/(f + 1)$ becomes more and more important with increasing numbers of levels, a point at which further increases in the number of levels is no longer justified exists. That is, for every experiment in which the null hypothesis is false, there must exist an optimal number of levels at which ${}_1I_a$ is a minimum. Fewer or more levels than this optimal number will result in less precision in the evaluation of the treatment effects. Values of ${}_1I_a$ were computed for $t = 2, 5$; $N = 20, 30, 50, 70, 100, 150$ to determine the number of levels at which this minimum is reached and the value of ${}_1I_a$ at this point. These data are presented in Tables 3 and 4. For comparative purposes the

TABLE 3

Factorial Design: Optimal Number of Levels for Selected Experimental Conditions, Assuming Levels Defined by Equal Proportions of the Population

ρ	t	N					
		20	30	50	70	100	150
.2	2	2	3	4	5	7	9
	5	1	2	2	3	4	6
.4	2	3	4	6	9	13	17
	5	2	3	4	5	7	10
.6	2	4	6	9	13	17	25**
	5	2	3	5	7	9	14
.8	2	5*	7*	12*	17	23	25**
	5	2*	3*	5*	7*	10*	15*

* Limit imposed by the requirement $N/th \geq 2$.

** Slight improvement possible with more than 25 levels.

TABLE 4

Factorial Design (with Optimum Number of Levels) and Completely Randomized Design: Values of ${}_1I_a$ for Selected Experimental Conditions*

ρ	t	N					
		20	30	50	70	100	150
.2	2	1.134 (1.151)	1.089 (1.114)	1.053 (1.084)	1.037 (1.072)	1.026 (1.063)	1.017 (1.056)
	5	1.172 (1.172)	1.112 (1.122)	1.065 (1.087)	1.045 (1.073)	1.031 (1.063)	1.020 (1.056)
.4	2	1.178 (1.316)	1.116 (1.273)	1.067 (1.239)	1.047 (1.225)	1.032 (1.214)	1.021 (1.206)
	5	1.263 (1.339)	1.169 (1.282)	1.093 (1.242)	1.064 (1.227)	1.043 (1.215)	1.028 (1.207)
.6	2	1.244 (1.727)	1.157 (1.670)	1.088 (1.626)	1.061 (1.608)	1.041 (1.594)	1.027 (1.583)
	5	1.423 (1.758)	1.256 (1.683)	1.139 (1.630)	1.095 (1.610)	1.063 (1.595)	1.040 (1.584)
.8	2	1.398 (3.070)	1.249 (2.970)	1.137 (2.891)	1.093 (2.858)	1.063 (2.834)	1.043 (2.815)
	5	1.944 (3.125)	1.539 (2.991)	1.274 (2.899)	1.179 (2.862)	1.115 (2.836)	1.071 (2.816)

*Value of I_a for completely randomized design appears in parentheses.

values of ${}_1I_a$ which hold for the completely randomized design have been included in parentheses in Table 4. For the completely randomized design

$$ {}_1I_a = \frac{(N - t + 3)}{(N - t + 1)(1 - \rho^2)}. $$

It may be noted in Table 3 that for several combinations of N, ρ, and t the optimal number of levels would not permit an equal, integral number of subjects per cell within a treatment. For example, with $N = 20$, $\rho = .4$, $t = 2$, the use of three levels would require either that one level include an extra subject, or that a total of 18 rather than 20 subjects be used. In the preparation of Table 4 the possible necessity of removing subjects was ignored, and I_a was computed from the full value of N. This procedure greatly facilitated the comparison of the precision of the various designs for selected values of N at the cost of only negligible error for a few of the selected conditions.

Within the scope of the values considered, the data in Tables 3 and 4 show several specific trends which experienced researchers may have already recognized. These relationships may be summarized as follows. The optimal number of levels tends to be larger for (i) larger values of ρ, (ii) larger numbers of experimental subjects, and (iii) smaller numbers of treatments. Each of these trends is consistent with the recognized effects of the size of the correlation between the criterion and control variables and of reductions in degrees of freedom on experimental precision. The optimal numbers of levels presented in Table 3 should serve as useful guides in the planning of two-factor experiments. Because changes in precision are relatively small as the number of levels approaches the optimal value, linear interpolations will yield suitably accurate estimates, except for interpolation along the scale of ρ. This interpolation should be made in terms of ρ^2.

Analysis of Covariance (*Design* 2)

The sample estimate of the variance of the difference between two adjusted means in an analysis of covariance is

$$ \text{var}\,(M'_i - M'_k) = E'_{YY}\left[\frac{2}{n} + \frac{(\bar{X}_1 - \bar{X}_2)^2}{f_e E_{XX}}\right]. \tag{7} $$

In this expression M'_i represents an adjusted treatment mean, E'_{YY} the adjusted mean square for error for the criterion measure, E_{XX} the mean square for error in X, and f_e the degrees of freedom for error in X. Finney [10] has recommended that when many pairs of treatments are to be tested, an average variance for error may be used. It may be computed as follows

$$ \text{ave var}\,(M'_i - M'_k) = \frac{2E'_{YY}}{n}\left[1 + \frac{T_{XX}}{f_e E_{XX}}\right]. \tag{8} $$

In this formula, T_{xx} represents the mean square for treatments on the control variable. Assuming a normal distribution for X, the second term within the brackets is an F ratio divided by f_e . Taking the expected value of this expression first with respect to Y and then with respect to X, then

$$\text{ave var } (M'_i - M'_k) = \frac{2\sigma_Y^2(1 - \rho^2)}{n}\left(1 + \frac{1}{f_e - 2}\right). \tag{9}$$

The value ${}_2I_t$ thus becomes

$${}_2I_t = 1 + \frac{1}{f_e - 2} = \frac{N - t - 1}{N - t - 2}. \tag{10}$$

Since the degrees of freedom for error in the analysis of adjusted scores is $N - t - 1$,

$${}_2I_a = \left(\frac{N - t - 1}{N - t - 2}\right)\left(\frac{N - t + 2}{N - t}\right). \tag{11}$$

TABLE 5

Analysis of Covariance: Values of ${}_2I_a$ for Selected Experimental Conditions

t	N					
	20	30	50	70	100	150
2	1.181	1.113	1.063	1.045	1.031	1.020
5	1.221	1.127	1.068	1.047	1.032	1.021

Values of ${}_2I_a$ for analyses of covariance have been tabulated in Table 5. From a comparison of these values with those in Table 4 note that for $\rho < .4$ the factorial approach results in approximately equal or greater precision than covariance; for $\rho \geq .6$ the advantage is in favor of covariance. For relatively high values of ρ and relatively small values of N the difference in precision is appreciable. This difference is mainly attributable to the fact that relatively small values of N do not permit the experimenter to employ a sufficiently large number of levels to exploit fully the value of the control variable. However, the marked superiority of covariance occurs for values of ρ which are rarely encountered in educational and psychological experiments. It may also be noted that for $\rho < .2$ and small values of N neither covariance nor the factorial design yields appreciably greater precision than a completely randomized design.

Differences (Design 3)

The method of differences consists of a completely randomized analysis applied to the measure $(Y - X)$. The variance of this measure is

$$\text{var}\,(X - Y) = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y \,. \tag{12}$$

Since the use of difference measures is generally restricted to instances in which X and Y may be regarded as parallel test forms or replicated measurements, assume $\sigma_X = \sigma_Y$. Thus

$$\begin{aligned} \text{var}\,(X - Y) &= 2\sigma_Y^2 - 2\rho\sigma_Y^2 \\ &= 2\sigma_Y^2(1 - \rho), \end{aligned} \tag{13}$$

and

$$\text{var}\,(M_i - M_k) = 4\sigma_Y^2(1 - \rho)/n. \tag{14}$$

Thus

$${}_3I_t = \frac{2}{1 + \rho}\,, \tag{15}$$

and

$${}_3I_a = \left(\frac{2}{1 + \rho}\right)\left(\frac{N - t + 3}{N - t}\right). \tag{16}$$

TABLE 6

Analysis of Variance of Differences: Values of ${}_3I_a$ for Selected Experimental Conditions

ρ	t	N					
		20	30	50	70	100	150
.2	2	1.842	1.782	1.735	1.715	1.700	1.689
	5	1.875	1.795	1.739	1.717	1.701	1.690
.4	2	1.579	1.527	1.487	1.470	1.457	1.448
	5	1.607	1.538	1.491	1.472	1.458	1.448
.6	2	1.382	1.336	1.301	1.286	1.275	1.267
	5	1.406	1.346	1.304	1.288	1.276	1.267
.8	2	1.228	1.188	1.156	1.143	1.134	1.126
	5	1.250	1.196	1.159	1.145	1.134	1.126

Values of ${}_3I_a$ for this design are tabulated in Table 6. Comparison of these with corresponding values for the factorial and covariance designs clearly indicates the lower precision of the difference approach. It is also indicated that unless a substantial correlation exists between the control and criterion variables the difference approach results in considerably lower precision than that yielded by the completely randomized design.

Discussion

The measures of comparative precision derived above are an important consideration, but not the sole consideration in the choice of experimental design. Serious attention must be given to the effect of possible departure from the assumptions on which the methods are based, the importance of design simplicity in the communication of results, and the extent to which valuable supplementary information may be derived from one or another of the designs. Several writers [5, 9, 11, 18] have pointed out that when two characteristics are to be controlled, blocking on one factor and covariance control on the other may be advantageous. This procedure, which is somewhat equivalent to multiple covariance, can yield valuable supplementary information and still retain the simplicity of analysis of simple covariance. However, the values of I_a derived by Cox strongly suggest that the combination can not be defended on the basis of the precision of treatment comparisons.

Most writers probably agree with Kempthorne ([16], p. 159) that dependence upon the accuracy of the assumed regression model constitutes a severe restriction on the usefulness of covariance techniques. The absence of any regression assumptions in the levels design, on the other hand, represents a considerable argument in its favor, especially in such instances as the number of degrees of freedom are fairly large and the difference in the precision of the designs is relatively insignificant. It would, in fact, seem justifiable to conclude from the data in Tables 4 and 5 that the less stringent assumptions of the factorial design more than compensate for the relatively small advantage in precision which may obtain for covariance. This is especially true in educational and psychological research in which the number of degrees of freedom for error is usually quite large and often relatively little is known about the form of the relationship between criterion and control measures. This presumes, of course, that the experimenter will include a number of levels which is reasonably close to the optimal value.

The more general applicability of the factorial design becomes apparent from the relationship between the homogeneity of regression from treatment population to treatment population and the phenomenon of interaction between treatments and levels in the two-factor design. It will be shown that heterogeneity of regression is equivalent to such an interaction, and the presence of either implies the other.

To prove this equivalence, first note that the phenomenon of inter-

action manifests itself by variation in the size of the treatments effects from at least one level of the control variable to one other. If, as before, $\bar{y}_{ij}$ represents the criterion mean of the subpopulation corresponding to level i and treatment j, the presence of interaction may be represented by the inequality

$$\bar{y}_{ij} - \bar{y}_{iv} \neq \bar{y}_{hj} - \bar{y}_{hv}$$

or

$$\bar{y}_{ij} - \bar{y}_{hj} \neq \bar{y}_{iv} - \bar{y}_{hv} , \tag{17}$$

where the subscripts i and h refer to levels, j and v to treatments.

If $f_j(x)$ and $f_v(x)$ represent the population regression functions, exclusive of any constant term, for treatments j and v, the four subpopulation means in these inequalities must satisfy the following equations.

$$\begin{aligned} \bar{y}_{ij} &= f_j(\bar{x}_i) - a_j , \qquad \bar{y}_{hj} = f_j(\bar{x}_h) - a_j ; \\ \bar{y}_{iv} &= f_v(\bar{x}_i) - a_v , \qquad \bar{y}_{hv} = f_v(\bar{x}_h) - a_v . \end{aligned} \tag{18}$$

The constants a_j and a_v represent the effects of treatments j and v, respectively, plus any constant term in the regression equations. Subtracting the second member of each pair from the first yields

$$\begin{aligned} \bar{y}_{ij} - \bar{y}_{hj} &= f_j(\bar{x}_i) - f_j(\bar{x}_h), \\ \bar{y}_{iv} - \bar{y}_{hv} &= f_v(\bar{x}_i) - f_v(\bar{x}_h). \end{aligned} \tag{19}$$

According to (17), however, the left sides of these equations are not equal. Therefore,

$$f_j(\bar{x}_i) - f_j(\bar{x}_h) \neq f_v(\bar{x}_i) - f_v(\bar{x}_h). \tag{20}$$

This inequality can hold only if $f_j(x)$ is not identical to $f_v(x)$. Thus it has been shown that the presence of interaction in the levels design implies heterogeneous regression in the covariance analysis.

To prove the converse, first note that heterogeneous regression means, by definition, nonidentical or nonparallel regression functions. That is, there exist at least two levels such that

$$f_j(\bar{x}_i) - f_v(\bar{x}_i) \neq f_j(\bar{x}_h) - f_v(\bar{x}_h). \tag{21}$$

But since $\bar{y}_{ij} = f_j(\bar{x}_i)$, it follows that

$$\bar{y}_{ij} - \bar{y}_{iv} \neq \bar{y}_{hj} - \bar{y}_{hv} . \tag{22}$$

Equation (22) defines an interaction between treatments j and v and levels h and i. Thus heterogeneous regression implies the presence of interaction.

If other assumptions are satisfied and if the combined levels subpopula-

tions constitute a meaningful experimental population, a valid and meaningful test of main effects of treatments may still be made in the factorial design against a within-cells estimate of error variance, even though an interaction exists. On the other hand, heterogeneous regression renders the covariance technique, as it is typically applied in educational and psychological research, somewhat invalid. The extent of this lack of validity has not been extensively investigated. If the usual covariance model is used, the effects would appear to be more serious than those of non-normality and heterogeneity of variance are to an analysis of variance. In cases of heterogeneity of regression, the obtained error variance would probably overestimate the true error variance, and thus increase the probability of retaining a false null hypothesis. Further research may indicate such violation of assumptions is no more serious than heterogeneity of variance or non-normality. However, no such conclusion seems warranted at this time.

There are available more general covariance models than that typically employed in educational and psychological research. These models may be applied to many cases of nonlinear and heterogeneous regression and allow a valid test of treatments effects, so long as the mathematical form of the regression equations may be specified on a priori grounds. The phenomenon of interaction, as manifested in the levels design, does not, therefore, entirely preclude a valid analysis by covariance techniques. However, the necessity of knowing the appropriate regression model represents a real restriction on the general applicability of covariance techniques. In most cases little is known about the fundamental nature of the measures used in psychological and educational research to make possible anything but the most tenuous assumption concerning the form of the relationship.

Then, too, the typical experimenter in psychology and education is not, as a rule, familiar with the more general covariance models. They are not illustrated or discussed in statistical texts intended for workers in these fields. Although heterogeneity of regression is equivalent to interaction between treatments and levels, few experimenters are accustomed to think of interaction in these terms. The process of interpreting and communicating experimental results is thus made more difficult.

Thus in such cases as the data on the control variable are available before the experiment is initiated, the most prudent design would almost certainly be that involving stratification rather than covariance. The latter technique, on the other hand, might be reserved for experiments in which stratification is not feasible.

It should be noted that marked interaction in the treatments-by-levels design will often be accompanied by heterogeneity of variance from treatment to treatment within the same level. If, for example, the interaction results from heterogeneous linear regression due to differences in ρ for the various treatments, the variance in the population at level i and treatment j will

not equal that at level i and treatment v. This may be seen from (3), which, with varying values of ρ, might be written

$$\sigma^2_{y_{ij}} = \sigma^2_Y\left[1 - \rho^2_i\left(1 - \frac{\sigma^2_{x_i}}{\sigma^2_X}\right)\right].$$

Varying values of ρ^2_i , coupled with constant values of $\sigma^2_{x_i}$, would result in some heterogeneity among the subpopulation variances at level i.

This degree of heterogeneity of variance would be only slightly more serious than that which has been demonstrated from level to level within any treatment. For example, if $\rho_j = .6$ and $\rho_v = .3$, the population at the third level in a six-level experiment such as that described in previous examples would have a variance of $.645\sigma^2_y$ in treatment j, $.911\sigma^2_y$ in treatment v. The ratio is less than 2 to 1, however, and this degree of heterogeneity would not, according to the findings of the investigators referred to earlier, seriously affect the validity of the F test. On the other hand, the effect of this degree of heterogeneity of regression on the validity of an analysis of covariance may well be more serious.

REFERENCES

[1] Anderson, R. L. and Bancroft, T. A. *Statistical theory in research.* New York: McGraw-Hill, 1952.

[2] Bartlett, M. S. The use of transformations. *Biometrics*, 1947, **3**, 39–52.

[3] Box, G. E. P. Some theorems on quadratic form applied in the study of analysis of variance problems, I. Effects of inequality of variance in the one-way classification. *Ann. math. Statist.*, 1954, **25**, 290–302.

[4] Cochran, W. G. Some consequences when the assumptions for analysis of variance are not satisfied. *Biometrics*, 1947, **3**, 22–28.

[5] Cochran, W. G. and Cox, G. M. *Experimental designs.* New York: Wiley, 1950.

[6] Cox, D. R. The use of a concomitant variable in selecting an experimental design. *Biometrika*, 1957, **44**, 150–158.

[7] Eisenhart, C. The assumptions underlying the analysis of variance. *Biometrics*, 1947, **3**, 1–21.

[8] Federer, W. T. *Experimental design.* New York: Macmillan, 1955.

[9] Federer, W. T. and Schlottfeldt, C. S. The use of covariance to control gradients in experiments. *Biometrics*, 1954, **10**, 282–290.

[10] Finney, D. J. Standard errors of yields adjusted for regression on an independent measurement. *Biometrics*, 1946, **2**, 53–55.

[11] Fisher, R. A. *Statistical methods for research workers.* (12th ed.) London: Oliver and Boyd, 1952.

[12] Fisher, R. A. *The design of experiments.* (5th ed.) London: Oliver and Boyd, 1949.

[13] Gourlay, N. Covariance analysis and its applications in psychological research. *Brit. J. statist. Psychol.*, 1953, **6**, 25-34.

[14] Graybill, F. Variance heterogeneity in a randomized block design. *Biometrics*, 1954, **10**, 516–520.

[15] Greenberg, B. G. Use of covariance and balancing in analytical surveys. *Amer. J. publ. Hlth*, 1953, **43**, 692–9.

[16] Kempthorne, O. *The design and analysis of experiments.* New York: Wiley, 1952.

[17] Lindquist, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1953.
[18] Lucas, H. L. Design and analysis of feeding experiments with milking dairy cattle. Raleigh, N. C.: Inst. Statist. Mimeo. Series No. 18, Univ. North Carolina, 1951.
[19] Norton, D. W. An empirical investigation of some effects of nonnormality and heterogeneity on the *F*-distribution. Unpublished doctoral dissertation, State Univ. Iowa, 1952.
[20] Outhwaite, A. D. and Rutherford, A. Covariance analysis as an alternative to stratification in the control of gradients. *Biometrics*, 1955, 11, 431–440.
[21] Snedecor, G. W. *Statistical methods.* (5th ed.) Ames, Iowa: Iowa State College Press, 1956.

Manuscript received 3/19/57
Revised manuscript received 3/6/58

PSYCHOMETRIKA—VOL. 23, NO. 4
DECEMBER, 1958

AN AXIOMATIC FORMULATION AND GENERALIZATION OF SUCCESSIVE INTERVALS SCALING*

ERNEST ADAMS
UNIVERSITY OF CALIFORNIA, BERKELEY
AND
SAMUEL MESSICK
EDUCATIONAL TESTING SERVICE

A formal set of axioms is presented for the method of successive intervals, and directly testable consequences of the scaling assumptions are derived. Then by a systematic modification of basic axioms the scaling model is generalized to non-normal stimulus distributions of both specified and unspecified form.

Thurstone's scaling models of successive intervals [7, 21] and paired comparisons [17, 24] have been severely criticized because of their dependence upon an apparently untestable assumption of normality. This objection was recently summarized by Stevens [22], who insisted that the procedure of using the variability of a psychological measure to equalize scale units "smacks of a kind of magic—a rope trick for climbing the hierarchy of scales. The rope in this case is the *assumption* that in the sample of individuals tested the trait in question has a canonical distribution, (e.g., 'normal') ⋯ . There are those who believe that the psychologists who make assumptions whose validity is beyond test are hoist with their own petard ⋯ ." Luce [13] has also viewed these models as part of an "extensive and unsightly literature which has been largely ignored by outsiders, who have correctly condemned the *ad hoc* nature of the assumptions."

Gulliksen [11], on the other hand, has explicitly discussed the testability of these models and has suggested alternative procedures for handling data which do not satisfy the checks. Empirical tests of the scaling theory were also mentioned or implied in several other accounts of the methods [e.g., 8, 9, 12, 15, 21, 25]. Criteria of goodness of fit have been presented [8, 18], which, if met by the data, would indicate satisfactory scaling within an acceptable error. Random errors and sampling fluctuations, as well as systematic deviations from scaling assumptions, are thereby evaluated by these

*This paper was written while the authors were attending the 1957 Social Science Research Council Summer Institute on Applications of Mathematics in Social Science. The research was supported in part by Stanford University under Contract NR 171-034 with Group Psychology Branch, Office of Naval Research, by Social Science Research Council, and by Educational Testing Service. The authors wish to thank Dr. Patrick Suppes for his interest and encouragement throughout the writing of the report and Dr. Harold Gulliksen for his helpful and instructive comments on the manuscript.

over-all internal consistency checks. However, tests of the scaling assumptions, and in particular the normality hypothesis, have not yet been explicitly derived in terms of the necessary and sufficient conditions required to satisfy the model. Recently Rozeboom and Jones [20] and Mosteller [16] have investigated the sensitivity of successive intervals and paired comparisons, respectively, to a normality requirement, indicating that departures from normality in the data are not too disruptive of scale values with respect to goodness of fit, but direct empirical consequences of the assumptions of the model were not specified as such.

The present axiomatic characterization of a well-established scaling model was attempted because of certain advantages which might accrue: (*a*) an ease of generalization that follows from a precise knowledge of formal properties by systematically modifying axioms, and (*b*) an ease in making comparisons between the properties of different models. The next section deals with the axioms for successive intervals and serves as the basis for the ensuing section, in which the model is generalized to non-normal stimulus distributions. One outcome of the following formalization which should again be highlighted is that the assumption of normality has directly verifiable consequences and should not be characterized as an untestable supposition.

Thurstone's Successive Intervals Scaling Model

The Experimental Method

In the method of successive intervals subjects are presented with a set of n stimuli and asked to sort them into k ordered categories with respect to some attribute. The proportion of times f_{si} that a given stimulus s is placed in category i is determined from the responses. If it is assumed that a category actually represents a certain interval of stimulus values for a subject, then the relative frequency with which a given stimulus is placed in a particular category should represent the probability that the subject estimates the stimulus value to lie within the interval corresponding to the category. This probability is in turn simply the area under the distribution curve inside the interval. So far scale values for the end points of the intervals are unknown, but if the observed probabilities for a given stimulus are taken to represent areas under a normal curve, then scale values may be obtained for both the category boundaries and the stimulus.

Scale values for interval boundaries are determined by this model, and interval widths are not assumed equal, as in the method of equal appearing intervals. Essentially equivalent procedures for obtaining successive intervals scale values have been presented by Saffir [21], Guilford [10], Mosier [15], Bishop [3], Attneave [2], Garner and Hake [9], Edwards [7], Burros [5], and Rimoldi [19]. The basic rationale of the method had been previously outlined by Thurstone in his absolute scaling of educational tests [23, 26]. Gulliksen

[12], Diederich, Messick, and Tucker [6], and Bock [4] have described least square solutions for successive intervals, and Rozeboom and Jones [20] presented a derivation for scale values which utilized weights to minimize sampling errors. Most of these papers contain the notion that the assumption of normality can be checked by considering more than one stimulus. Although one distribution of relative frequencies can always be converted to a normal curve, it is by no means always possible to normalize simultaneously all of the stimulus distributions, allowing unequal means and variances, on the same base line. The specification of exact conditions under which this is possible will now be attempted. In all that follows, the problem of sampling fluctuations is largely ignored, and the model is presented for the errorless case.

The Formal Model

The set of stimuli, denoted S, has elements $r, s, u, v, \cdots$. There is no limit upon the admissible number of stimuli, although for the purpose of testing the model, S must have at least two members. For each stimulus s in S, and each category $i = 1, 2, \cdots, k$, the relative frequency $f_{s,i}$ with which stimulus s is placed in category i is given. Formally f is a function from the Cartesian product of $S \times \{1, 2, \cdots, k\}$ to the real numbers. More specifically, it will be the case that for each s in S, f_s will be a probability distribution over the set $\{1, 2, \cdots, k\}$. For the sake of an explicit statement of the assumptions of the model, this fact will appear as an axiom, although it must be satisfied by virtue of the method of determining the values of $f_{s,i}$.

Axiom 1. f is a function mapping $S \times \{1, \cdots, k\}$ into the real numbers such that for each s in S, f_s is a probability distribution over $\{1, \cdots, k\}$; i.e., for each s in S and $i = 1, \cdots, k$, $0 \leq f_{s,i} \leq 1$ and $\sum_{i=1}^{k} f_{s,i} = 1$.

The set S and the function f constitute the *observables* of the model. Two more concepts which are not directly observed remain to be introduced. The first of these is a set of numbers $t_1, \cdots, t_{(k-1)}$, which are the end points of the intervals corresponding to the categories. It is assumed that these intervals are adjacent and that they cover the entire real line. Formally, it will simply be assumed that $t_1, \cdots, t_{(k-1)}$ are an increasing series of real numbers.

Axiom 2. Interval boundaries $t_1, \cdots, t_{(k-1)}$ are real numbers, and for $i = 2, \cdots, (k-1)$, $t_{(i-1)} \leq t_i$.

Finally, the distribution corresponding to each stimulus s in S is represented by a normal distribution function N_s.

Axiom 3. N is a function mapping S into normal distribution functions over the real line.

Axioms 1–3 do not state fully the mathematical properties required for

the set S, the numbers $t_1, \cdots, t_{(k-1)}$, and the functions N_s. In the interests of completeness, these will be stated in the following Axiom 0, which for formal purposes should be referred to instead of Axioms 1–3.

AXIOM 0. S is a non-empty set. k is a positive integer. f is a function mapping $S \times \{1, \cdots, k\}$ into the closed interval [0, 1], such that for each s in S, $\sum_{i=1}^{k} f_{s,i} = 1$. For $i = 1, \cdots, (k-1)$, t_i is a real number, and for $i = 1, \cdots, (k-2)$, $t_i \leq t_{i+1}$. N is a function mapping S into the set of normal distribution functions over the real numbers.

Axioms 2 and 3 state only the set-theoretical character of the elements t_i and N_s, and have no intuitive empirical content. The central hypothesis of the theory states the connection between the observed relative frequencies $f_{s,i}$ and the assumed underlying distributions N_s.

AXIOM 4. (Fundamental hypothesis) For each s in S and $i = 1, \cdots, k$,

$$f_{s,i} = \int_{t_{i-1}}^{t_i} N_s(\alpha)\, d\alpha.$$

(Note that if $i = 1$, $t_{(i-1)}$ is set equal to $-\infty$, and if $i = k$, $t_i = \infty$.)

Axioms 1–4 state the formal assumptions of the theory although, because the fundamental hypothesis (Axiom 4) involves the unobservables N_s and t_i, it is not directly testable in these terms. The question of testing the model will be discussed in the next section. Scale values for the stimuli have not yet been introduced. These are defined to be equal to the means of the distributions N_s, and hence are easily derived. The function v will represent the scale values of the stimuli.

DEFINITION 1. v is the function mapping S into the real numbers such that for each s in S, v_s is the mean of N_s; i.e.,

$$v_s = \int_{-\infty}^{\infty} \alpha N_s(\alpha)\, d\alpha.$$

Testing the Model

The model will be said to fit exactly if all of the testable consequences of Axioms 1–4 are verified. Testable consequences of these axioms will be those consequences which are formulated solely in terms of the observable concepts S and f, or of concepts which are definable in terms of S and f. If no further assumptions are made about an independent determination of $t_1, \cdots, t_{(k-1)}$ and N, then the testable consequences are just those which follow about f and S from the assumption that there *exist* numbers $t_1, \cdots, t_{(k-1)}$ and functions N_s which satisfy Axioms 1–4. In this model, it is possible to give an exhaustive description of the testable consequences; hence this theory is axiomatizable in the sense that it is possible to formulate observable conditions which are necessary and sufficient to insure the existence

of the numbers t_i and functions N_s. The derivation of these conditions will proceed by stages.

Let $p_{s,i}$ be the cumulative distribution of the function f for stimulus s and interval i.

DEFINITION 2. For each s in S and $i = 1, \cdots, k$,

$$p_{s,i} = \sum_{j=1}^{i} f_{s,j} .$$

It follows from this definition and Axiom 4 that for each s in S and $i = 1, \cdots, k$,

$$p_{s,i} = \int_{-\infty}^{t_i} N_s(\alpha)\, d\alpha. \tag{1}$$

Using the table for the cumulative distribution of the normal curve with zero mean and unit variance, the numbers $z_{s,i}$ may be determined such that

$$p_{s,i} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_{s,i}} e^{-1/2x^2}\, dx. \tag{2}$$

(Note that for $i = k$, $z_{s,i}$ will be infinite.) N_s is a normal distribution function and must have the form:

$$N_s(\alpha) = \frac{1}{\sigma_s \sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_s^2}(\alpha - v_s)^2\right], \tag{3}$$

where σ_s^2 is the variance of N_s about its mean v_s. Equations (1), (2), and (3) yield the conclusion that for each s in S and $i = 1, \cdots, k$,

$$z_{s,i} = (t_i - v_s)/\sigma_s . \tag{4}$$

In (4) the numbers $z_{s,i}$ on the left are known transformations of the observed proportions $f_{s,i}$, while the numbers t_i, v_s and σ_s are unknown. Suppose however that r is a fixed member of the class S of stimuli; it is possible to solve (4) for all the unknowns in terms of the known z's, and v_r and σ_r, the mean and standard deviation of the fixed stimulus r. These solutions are

$$t_i = \sigma_r z_{r,i} + v_r \quad \text{for} \quad i = 1, \cdots, (k-1); \tag{5}$$

$$\sigma_s = \sigma_r\left(\frac{z_{r,i} - z_{r,j}}{z_{s,i} - z_{s,j}}\right) \quad \text{for} \quad s \in S, \quad \text{and} \quad i \neq j; \tag{6}$$

$$v_s = \sigma_r\left[z_{r,i} - \left(\frac{z_{r,i} - z_{r,j}}{z_{s,i} - z_{s,j}}\right) z_{s,i}\right] + v_r . \tag{7}$$

The necessary and sufficient condition that the system of equations (4) have a solution, and hence that t_i, v_s and σ_s be determinable using (5), (6),

and (7), is that all $z_{s,i}$ be linear functions of each other in the following sense. For all r and s in S, there exist real numbers $a_{r,s}$ and $b_{r,s}$ such that for each $i = 1, \cdots, k$,

$$z_{s,i} = a_{r,s}z_{r,i} + b_{r,s} . \tag{8}$$

The required numbers $a_{r,s}$ and $b_{r,s}$ exist if and only if for each r and s, the ratio

$$\frac{z_{r,i} - z_{r,j}}{z_{s,i} - z_{s,j}} = a_{s,r} = \frac{1}{a_{r,s}} \tag{9}$$

is independent of i and j.

If constants $a_{r,s}$ and $b_{r,s}$ satisfying (8) exist, then they are related to the scale values v_r and the standard deviations σ_r in a simple way. For each r, s in S,

$$a_{r,s} = \sigma_r/\sigma_s , \tag{10}$$

and

$$b_{r,s} = (v_r - v_s)/\sigma_s . \tag{11}$$

Clearly the arbitrary choice of the constants v_r and σ_r in (5), (6), and (7) represents the arbitrary choice of origin and unit in the scale. Since scale values of t_i and v_s are uniquely determined once v_r and σ_r are chosen, the scale values are unique up to a linear transformation; i.e., an interval scale of measurement has been determined. It should be noted that this model does not require equality of standard deviations (or what Thurstone has called discriminal dispersions [25]) but provides for their determination from the data by equation (6). This adds powerful flexibility in its possible applications.

It remains only to make a remark about the necessary and sufficient condition which a set of observed relative frequencies $f_{s,i}$ must fulfill in order to satisfy the model. This necessary and sufficient condition is simply that the numbers $z_{s,i}$, which are defined in terms of the observed relative frequencies, be linearly related as expressed in (8). This can be determined by seeing if the ratios computed from (9) are independent of i and j, or by evaluating for all s, r the linearity of the plots of $z_{s,i}$ against $z_{r,i}$. Hence for this model there is a simple decision procedure for determining whether or not a given set of errorless data fits.

If $z_{s,i}$ and $z_{r,i}$ are found to be linearly related for all s, r in S, the assumptions of the scaling model are verified for that data. If the z's are not linearly related, then assumptions have been violated. For example, the normal curve may not be an appropriate distribution function for the stimuli and some other function might yield a better fit [cf. 11, 12]. Or perhaps the responses cannot be summarized unidimensionally in terms of projections on the real line representing the attribute [11]. If the stimuli are actually distributed in a

multidimensional space, then judgments of projections on one of the attributes may be differentially distorted by the presence of variations in other dimensions. This does not mean that stimuli varying in several dimensions may not be scaled satisfactorily by the method of successive intervals, but rather that if the model does not fit, such distortion effects might be operating. A multidimensional scaling model [14] might prove more appropriate in such cases.

In practice the set of points $(z_{r,i}, z_{s,i})$ for $i = 2, \cdots, (k - 1)$ will never exactly fit the straight line of (8) but will fluctuate about it. It remains to be decided whether this fluctuation represents systematic departure from the model or error variance. In the absence of a statistical test for linearity, the decision is not precise, although the linearity of the plots may still be evaluated, even if only by eye. One approach is to fit the obtained points to a straight line by the method of least squares and then evaluate the size of the obtained minimum error [4, 6, 12]. In any event, the test of the model is exact in the errorless case, and the incorporation of a suitable sampling theory would provide decision criteria for direct experimental applications.

A Generalization of the Successive Intervals Model

The successive intervals model discussed in the previous section can be generalized in a number of ways. One generalization, treated in detail by Torgerson [27], considers each interval boundary t_i to be the mean of a subjective distribution with positive variance. Another approach toward generalizing the model is to weaken the requirement of normal distributions of stimulus scale values. Formally, this generalization amounts to enlarging the class of admissible distribution functions. Instead of specifying exactly which distribution functions are allowed in the generalization, assume an arbitrary set ψ of distributions over the real line, to which it is required that the stimulus distributions belong. In formalizing the model, ψ is characterized simply as a set of distribution functions over the real line. Axiom 3 may be replaced by a new axiom specifying the nature of the class ψ and stating that C is a function mapping S into elements of ψ; i.e., for each s in S, C_s (interpreted as the distribution of the stimulus s) is a member of ψ.

One final assumption about the class ψ needs to be added: namely, if ψ contains a distribution function C, then it must contain all *linear transformations* of C. A linear transformation of a distribution function C is defined as any other distribution function C′ which can be obtained from C by a shift of origin and a scale transformation of the horizontal axis. A stretch along the horizontal axis must be compensated for by a contraction on the vertical axis in order that the transformed function also be a probability density function. Algebraically, these transformations have the following form. Let D and D' be distribution functions, then D' is a linear transformation

of D if there exists a positive real number a and a real number b such that for all x,

$$D'(x) = aD(ax + b).$$

This is not truly a linear transformation because of multiplication by a on the ordinate, but for lack of a better term this phrase is used. The reason for requiring that the class ψ of distribution functions be closed under linear transformations is to insure that in any determination of stimulus scale values it will be possible to convert them by a linear transformation into another admissible set of scale values; i.e., the stimulus values obtained are to form an interval scale. If the set ψ is not closed under linear transformations, in general it will not be possible to alter the scale by an arbitrary linear transformation.

AXIOM 3′. ψ is a set of distribution functions over the real numbers, and C is a function mapping S into ψ. For all D in ψ, if a is a positive real number and b is a real number, then the function D' such that for all x,

$$D'(x) = aD(ax + b)$$

is a member of ψ.

It is to be observed that the set of normal distributions has the required property of being closed under linear transformations. This set is in fact a minimal class of this type, in the sense that all normal distribution functions can be generated from a single normal distribution function by linear transformations.

Finally, Axiom 4 is replaced by an obvious generalization which specifies the connection between the observed $f_{s,i}$, the distribution functions C_s, and the interval end points t_i.

AXIOM 4′. For each s in S and $i = 1, \cdots, k$,

$$f_{s,i} = \int_{t_{i-1}}^{t_i} C_s(x)\, dx.$$

(Here again $t_0 = -\infty$ and $t_k = \infty$.) The stimulus values are defined as before to be the means of the distribution functions C_s.

DEFINITION 1′. v is the function mapping S into the real numbers such that for each s in S, v_s is the mean of C_s, i.e.,

$$v_s = \int_{-\infty}^{\infty} xC_s(x)\, dx.$$

The problem now is to specify the class of admissible distribution functions ψ. Each specification of this class amounts to a theory about the underlying stimulus distributions. If the hypothesis of normality is altered or

weakened, what assumptions can replace it? Omitting any assumption about the form of the distribution functions would amount to letting ψ be the set of all distribution functions over real numbers. If no assumption whatever is made about the forms of C_s, then the theory is very weak. Every set of data will fit the theory, and the scale values of t_i can be determined only on an ordinal scale. It is always possible to determine distribution functions C_s satisfying Axiom 4′ for arbitrarily specified t_i. To show this it is only necessary to construct them in accordance with the following definition.

$$C_s(x) = \begin{cases} \dfrac{f_{s,i}}{t_i - t_{i-1}}, & i - 1 < x < i, \quad i = 1, \cdots, k, \\ 0 & \text{otherwise.} \end{cases}$$

Non-normal Distributions of Specified Form

It is clearly necessary to make some restrictions on ψ if the scale values are to be determined uniquely up to a linear transformation. It will next be shown that any minimal class of distribution functions, in the sense of a class all of whose members are generated from a single member by linear transformations, has the desired property of generating a linear scale of stimulus values when the model fits. For the present assume that ψ is a minimal class of distribution functions.

ASSUMPTION 1. There exists a distribution function D such that for all distribution functions D' in ψ there exists a positive real number a and a real number b such that for all x,

$$D'(x) = aD(ax + b).$$

To show that if Assumption 1 is satisfied the scale values are obtained on an interval scale, we proceed as follows. Axiom 3′ and Assumption 1 imply that for all s in S, there exists a positive real number a_s and a real number b_s such that for all x,

$$C_s(x) = a_s D(a_s x + b_s), \tag{12}$$

where the function D on the right side of (12) is a fixed function of some specified form linearly related to all the functions D' in ψ. According to Axiom 4′, then, for each s in S, and $i = 1, \cdots, k$,

$$f_{s,i} = \int_{t_{i-1}}^{t_i} a_s D(a_s x + b_s)\, dx. \tag{13}$$

If π is the cumulative distribution corresponding to D, and the cumulative distributions $p_{s,i}$ are defined as before, then

$$\begin{aligned} p_{s,i} &= \int_{-\infty}^{t_i} a_s D(a_s x + b_s)\, dx \\ &= \pi(a_s t_i + b_s). \end{aligned} \tag{14}$$

Assuming that the function π is strictly monotone increasing, then, knowing the form of function D, it is possible to determine uniquely the numbers $z_{s,i}$ such that for each s in S and $i = 1, \cdots, k$,

$$p_{s,i} = \pi(z_{s,i}). \tag{15}$$

Equations (14) and (15) imply immediately that

$$z_{s,i} = a_s t_i + b_s \tag{16}$$

for all s in S and $i = 1, \cdots, k$. It is clear from (15) why it is necessary to assume that π is strictly monotone increasing. If it were not, there would not in general be a unique $z_{s,i}$ determined by (15); hence the scale values based on $z_{s,i}$ would not be unique. It is also seen that (4), relating $z_{s,i}$ to t_i, v_s and σ_s in the normal distribution model, is simply a particular case of (16) here. The connection between a_s, b_s and σ_s and v_s is

$$\sigma_s = 1/a_s, \qquad v_s = -b_s/a_s.$$

In (15), as in the corresponding set of equations obtained from the normality assumption, the numbers on the left are known, and the numbers on the right are unknown. As before, if two numbers a_r and b_r are arbitrarily determined for a fixed stimulus r, then the t_i are uniquely determined by the following equation.

$$t_i = (z_{r,i} - b_r)/a_r, \qquad i = 1, \cdots, k. \tag{17}$$

The scale values for the stimuli, however, cannot be directly determined from the coefficients $z_{s,i}$, a_r and b_r without first specifying the mean m of the basic distribution D. If m is the mean of D, then v_s, which was defined as the mean of C_s, is determined by

$$v_s = (m - b_s)/a_s. \tag{18}$$

Both the a_s and the b_s in (17) can be determined in terms of $z_{s,i}$, a_r and b_r, (19) and (20); hence v_s is immediately determinable in terms of just these quantities by (18).

$$a_s = a_r \frac{z_{s,i} - z_{s,j}}{z_{r,i} - z_{r,j}}, \tag{19}$$

$$b_s = z_{s,i} - \left(\frac{z_{s,i} - z_{s,j}}{z_{r,i} - z_{r,j}}\right)(z_{r,i} - b_r). \tag{20}$$

It is clear then that the scale values t_i and v_s are determined up to a linear transformation. Furthermore, necessary and sufficient conditions that a set of data fit the model are simply that the ratios of differences in z's on the right in (19) be independent of i and j; i.e., that the z's be linearly related.

The Forms of the Distributions Unspecified

A final generalization to be considered is one in which Assumption 1 holds, but where the form of the generating function D is not specified; i.e., it is assumed that the underlying distributions all belong to one minimal class, but that the class can be generated by any distribution function D. Interestingly enough, in this case it is still possible to test the model and to obtain more than ordinal information about the scale values. If it is assumed that the stimulus distributions all belong to one minimal family generated by a function D, but D is unknown, all of the deductions up through (14) go through, although in this case the function π is also unknown. Now, of course, it is impossible to discover the numbers $z_{s,i}$ by solving (15), but if it is postulated that the function π is strictly monotone increasing, it is still possible to obtain some information about the numbers $(a_s t_i + b_s)$. Since π is a cumulative distribution it is monotone increasing; however, it will only be strictly monotone increasing in case the distribution function D is never zero. This assumption is made explicit in Assumption 2.

ASSUMPTION 2. For all x, $D(x) > 0$.

Now, if π is strictly monotone increasing, then it follows that $\pi(x) \geq \pi(y)$ if and only if $x \geq y$. If (14) holds, then it will be the case that for all r, s in S and $i, j = 1, \cdots, k$,

$$p_{s,i} \geq p_{r,j} \quad \text{if and only if} \quad a_s t_i + b_s \geq a_r t_j + b_r \,. \tag{21}$$

Therefore from an ordering on the numbers $p_{s,i}$ one can obtain a system of inequalities involving the constants a_s, b_s, and t_i. If it is further specified (as is required for the conditions of the problem) that $a_s > 0$ for all S, then this set of inequalities will not in general have a solution.

However, whether or not a set of data fits the model may still be determined. The necessary and sufficient condition for fit is that there exist numbers a_s, t_i and b_s (where $a_s > 0$) satisfying the system of inequalities (21). If this set of inequalities has a solution, then the interval boundaries may be taken to be the t_i satisfying (21). To determine the scale values of the stimuli it is first necessary to construct a distribution function which can represent the data. This is done in the following way. A differentiable monotone increasing function $\pi(x)$ is constructed by connecting the discrete set of points

$$\pi(a_s t_i + b_s) = p_{s,i}$$

with any smooth, strictly monotone increasing curve. If, as is usual, there is only a finite number of stimuli, then such a curve can always be constructed. Finally, the distribution function D is defined by

$$D(x) = \frac{d}{dx}\pi(x). \tag{22}$$

Then, if the mean of the distribution D is m, the values v_s of the stimuli are determined by (18), $v_s = (m - b_s)/a_s$. As far as the determination of the v_s is concerned, it can be seen that they depend solely on the previously determined a and b and on the mean m, which can be regarded as an additional arbitrary constant in the determination of the v_s .

The remaining point of discussion for this model is the determination of the degree of uniqueness of the scale values. Finding the set of all possible solutions to the inequalities (21) presents, in general, extreme difficulty. One thing that can be simply determined is the class of what might be called the *universal transformations* of the solutions of the system of inequalities. A universal transformation is one which, applied to a solution of any set of inequalities, yields another solution to the same set of inequalities. By noting a close connection between the theory of the inequalities (21) and a two-dimensional affine geometry with a distinguished set of horizontal and vertical lines, it can be shown [1] that the class of universal transformations for this model is a subset of the affine transformations. The universal transformations of the interval boundaries t_i are the linear ones, and of the a_s are multiplications by a positive constant. The b_s also are determined up to a linear transformation, and hence so are the scale values v_s (although the additional arbitrary constant m also enters into their determination).

There is also an interesting special case in which, even though there is only a finite number of observations, the scale values of the t_i are determined up to a linear transformation. This might be called the special case of equal intervals, in which differences in successive t_i are all the same. If, for example, there exist stimuli with such relations among corresponding p's as $p_{x,i} = p_{y,i+1} = p_{z,i+2}$, $p_{x,i+1} = p_{y,i+2}$, $p_{y,i} = p_{z,i+1}$, etc., it is possible to determine that successive intervals are equal [1].

The fact that scale values obtained in this model, at least under certain circumstances, are unique up to a linear transformation has two interesting consequences for the original successive intervals model based on the normality hypothesis. (i) If in the errorless case the original model fits, then *no other successive intervals model which assumes a different form for the distribution functions will fit.* The reason for this is that the forms of the distribution functions (or the cumulative distributions) are determined by the values of $p_{s,i}$ lying above the point t_i . Hence, if the t_i are determined up to linear transformation, so are the curves $p_{s,i}$. (ii) Where the normality assumption does not fit the data it is theoretically possible to use the present generalization to obtain a scale. Then the deviation of the scale values from those obtained under a normality requirement can be evaluated. This, at least in principle, provides a second kind of goodness of fit besides the usual least squares regression methods employed where the data do not exactly fit the Thurstone model.

REFERENCES

[1] Adams, E. and Messick, S. An axiomatization of Thurstone's successive intervals and paired comparisons scaling models. Stanford, Calif.: Stanford Univ., Applied Mathematics and Statistics Laboratory, ONR Technical Report No. 12, 1957.

[2] Attneave, F. A method of graded dichotomies for the scaling of judgments. *Psychol. Rev.*, 1949, **56**, 334–340.

[3] Bishop, R. Points of neutrality in social attitudes of delinquents and non-delinquents. *Psychometrika*, 1940, **5**, 35–45.

[4] Bock, R. D. Note on the least squares solution for the method of successive categories. *Psychometrika*, 1957, **22**, 231–240.

[5] Burros, R. H. The estimation of the discriminal dispersion in the method of successive intervals. *Psychometrika*, 1955, **20**, 299–305.

[6] Diederich, G., Messick, S., and Tucker, L. R. A general least squares solution for successive intervals. *Psychometrika*, 1957, **22**, 159–173.

[7] Edwards, A. L. The scaling of stimuli by the method of successive intervals. *J. appl. Psychol.*, 1952, **36**, 118–122.

[8] Edwards, A. L. and Thurstone, L. L. An internal consistency check for scale values determined by the method of successive intervals. *Psychometrika*, 1952, **17**, 169–180.

[9] Garner, W. R. and Hake, H. W. The amount of information in absolute judgments. *Psychol. Rev.*, 1951, **58**, 446–459.

[10] Guilford, J. P. The computation of psychological values from judgments in absolute categories. *J. exp. Psychol.*, 1938, **22**, 32–42.

[11] Gulliksen, H. Paired comparisons and the logic of measurement. *Psychol. Rev.*, 1946, **53**, 199–213.

[12] Gulliksen, H. A least squares solution for successive intervals assuming unequal standard deviations. *Psychometrika*, 1954, **19**, 117–139.

[13] Luce, R. D. A theory of individual choice behavior. Bureau Appl. Soc. Res., Columbia Univ., 1957. (Mimeo.)

[14] Messick, S. Some recent theoretical developments in multidimensional scaling. *Educ. psychol. Measmt*, 1956, **16**, 82–100.

[15] Mosier, C. I. A modification of the method of successive intervals. *Psychometrika*, 1940, **5**, 101–107.

[16] Mosteller, F. Some miscellaneous contributions to scale theory: Remarks on the method of paired comparisons. Cambridge: Harvard Univ., Lab. Soc. Relations, Report No. 10. Ch. III.

[17] Mosteller, F. Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 1951, **16**, 3–9.

[18] Mosteller, F. Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. *Psychometrika*, 1951, **16**, 207–218.

[19] Rimoldi, H. J. A. and Hormaeche, M. The law of comparative judgment in the successive intervals and graphic rating scale methods. *Psychometrika*, 1955, **20**, 307–318.

[20] Rozeboom, W. W. and Jones, L. V. The validity of the successive intervals method of psychometric scaling. *Psychometrika*, 1956, **21**, 165–183.

[21] Saffir, M. A comparative study of scales constructed by three psychophysical methods. *Psychometrika*, 1937, **2**, 179–198.

[22] Stevens, S. S. Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951.

[23] Thurstone, L. L. A method of scaling psychological and educational tests. *J. educ. Psychol.*, 1925, **16**, 433–451.

[24] Thurstone, L. L. Psychophysical analysis. *Amer. J. Psychol.*, 1927, **38**, 368–389.
[25] Thurstone, L. L. A law of comparative judgment. *Psychol. Rev.*, 1927, **34**, 424–432.
[26] Thurstone, L. L. The unit of measurement in educational scales. *J. educ. Psychol.*, 1927, **18**, 505–524.
[27] Torgerson, W. S. A law of categorical judgment. In L. S. Clark (Ed.), *Consumer behavior*. New York: New York Univ. Press, 1954.

Manuscript received 2/12/58
Revised manuscript received 4/21/58

PSYCHOMETRIKA—VOL. 23, NO. 4
DECEMBER, 1958

THE SINGLE LATIN SQUARE DESIGN IN PSYCHOLOGICAL RESEARCH

JOHN GAITO

AIR CREW EQUIPMENT LABORATORY, PHILADELPHIA†

The expected value of mean square concept is used to determine the effects of the presence of interactions in the single Latin square design on *F* tests. The results indicate that as the number of random effects included in the experiment increase, more *F* tests are unbiased, and that some of these are valid *F* tests. However, when *F* test bias does occur it is almost always of a negative nature so that the conclusions stated are conservative ones. Positive *F* test bias may occur when the triple interaction is extant and when zero or one random variate is included in the experiment.

Psychologists, mathematical statisticians, and others who utilize statistical techniques have made extensive use of Latin square designs. Generally, it has been assumed that interactions must be nonextant for results to be adequate. Thus some have raised questions concerning the suitability of these techniques for psychological research in which one of the variables is subjects, since interactions involving subjects frequently occur. For example, McNemar [6] maintains that if the interactions were not zero, obtained *F* values would not follow the *F* distribution and too many significant results would occur, a positive *F* test bias. By McNemar's argument, nonzero interaction results in a residual term larger than the ordinary error component, but the combination of the two sources yields a residual smaller than the interaction that should be used as the denominator for the *F* test. Lindquist [5] also has maintained that the single Latin square design will seldom be useful in educational and psychological research. He argues that each main effect is confounded with the interaction of the other two factors and with the triple interaction; Lindquist also stresses the ambiguous character of the residual.

Through use of expected values of mean squares, the adequacy of the Latin square designs may be better evaluated. Thus Gourlay [2], using a variance component analysis, indicates that for a valid application of the Latin square techniques interactions do not always have to be zero. Furthermore, contrary to McNemar's assertion, he found that too few significant *F* values might result, a negative *F* test bias.

Gourlay investigated this problem in reference to two main types of interactions that occur in psychology.

†Now at Wilkes College, Wilkes Barre, Pa.

(i) Each individual or unit receives only one of several treatments and is represented by one measurement in the data. In this case interaction is between main effects.

(ii) Repeated measurements are made on the same individuals or groups. In this case earlier measurements may interact with those that follow.

However, a more general and instructive procedure would be to determine the components of variance included within each mean square under four conditions: zero, one, two, and three random variates. The first condition corresponds to the fixed variate model, the last to the random variate model, and the others to the mixed model. This procedure would exhaust all possibilities in the single Latin square design and would permit an evaluation of the behavior of each test of significance.

In obtaining the expected value of mean square components, the procedure of Anderson and Bancroft [1], Greenwood [3], Kempthorne [4], and Tukey [9] or that used by McNemar [7] and Mood [8] might be employed. The two procedures differ in the components included in the expected value of mean square when random and fixed effects occur. The first procedure differs from the second by excluding from the random effect the variance due to the interaction of the fixed and random variates. The former procedure is favored by several intuitive arguments and will be used in this paper.

In a complete factorial design the rules for obtaining expected components may be stated simply [see 3]. The expected value of mean square for any source of variation is σ_e^2 (variance due to error) plus the σ^2 term having exactly the subscripts corresponding to the letters describing the source of variation. It further includes all σ^2 terms which have these same subscripts, providing the remaining subscripts all refer to random effects. These rules may be extended to the Latin square design with few modifications. The expected value of mean square for each main effect would contain the triple interaction and the double interaction of the other two factors, in addition to those required as stated above. Likewise, the residual would contain σ_e^2 and all interaction variances. However, the coefficients of all components except σ_e^2 would not be the same as those in the complete factorial design, since all levels of the three variables are not included in the Latin square design. These coefficients have been indicated in a recent paper by Wilk and Kempthorne [10], which presents a generalized derivation for Latin square designs and a special case where only fixed effects are involved. The coefficients and components are presented in Tables 1–4.

This paper will attempt to demonstrate the possible defects inherent in the single Latin square tests of significance when interactions occur under the four conditions mentioned above. If the variable in the experiment includes all levels in the population of interest, then the effect is considered

fixed. If the variable represents a random sampling from a population and the sample includes only a few levels of this variable, the effect is considered random. For each of the four conditions the effects of the presence of one or more interactions on tests of significance will be considered.

Zero Random Variate Model

In this model the levels of the three effects include the total population of levels and are considered as fixed. The paradigm for this model is shown in Table 1. The coefficients are either zero or one for all components except for the triple interaction and the three single effects. The coefficients for σ^2_{abc} in A, B, and C are the same, $t - (2/t)$, but differ from that of the residual, $t - (3/t)$. Each of the three single effects have a coefficient of t, the number of levels of each of the three main effects included in the experiment. The procedure below can be followed by the reader by eliminating components that do not occur in each case.

TABLE 1

Zero Random Variates Model

Variates	EMS (Coefficients of)							
	σ_e^2	σ_{abc}^2	σ_{bc}^2	σ_{ac}^2	σ_{ab}^2	σ_c^2	σ_b^2	σ_a^2
A	1	(t-2)/t	1	0	0	0	0	t
B	1	(t-2)/t	0	1	0	0	t	
C	1	(t-2)/t	0	0	1	t		
Residual	1	(t-3)/t	1	1	1			

Note - t is the number of levels of A, B, and C in the experiment.

Case 1. *One Double Interaction*

The F tests of both variates included in the interaction are negatively biased inasmuch as an inflation of the denominator occurs, resulting in an increase in the probability of a Type II error. The residual contains variance due to error and variance due to interaction. The mean squares of two of the main effects include variance due to error and variance due to interaction. The interaction variance would have to be included in both the numerator and denominator of the F test to achieve an unbiased result. The F test of the other variate is unbiased but is not a valid F test. To be a valid F ratio

more requirements than freedom from bias are imposed. The interactions contained in the mean squares of both the main effect and the residual must be random, normally distributed, and be a component that would be included in the mean square as indicated by the rule stated above. The F ratio in this instance is a ratio of two noncentral chi square statistics divided by their respective degrees of freedom, and the distribution depends upon the parameters of the fixed effect interaction. However, this F test does give a result above and beyond the interaction effect.

Case 2. *Triple Interaction*

The F test of each of the three variates contains a small positive bias inasmuch as the coefficient of σ^2_{abc} for residual is less than the coefficient of this term in A, B, and C.

Case 3. *Two Double Interactions*

No matter which two are present, all three F tests are negatively biased because only one interaction term is contained in A, B, and C while two are present in the residual.

Case 4. *Three Double Interactions*

All F tests contain negative bias, but the bias is greater than in Case 3 because three interaction terms are included in the residual and only one appears in each of the three main effects.

Case 5. *One Double Interaction and the Triple Interaction*

The results in this instance include both positive and negative biases. The F test of the variate not included in the first-order interaction has a small positive bias. The tests of the two variates included in the interaction will be biased, with the direction depending on the relative size of the variance due to the double and triple interactions. If the former is greater than $(1/t)$ times the latter, then negative bias occurs; if less than $(1/t)$, positive bias is present.

Case 6. *Two Double Interactions and the Triple Interaction*

F tests tend to be negatively biased. However, positive bias may occur. The F test of the variate which is included in both double interactions is negatively biased if $(1/t)\sigma^2_{abc}$ is less than the sum of the two double interaction variances. Positive bias occurs if the reverse is true. The F test of each of the other two effects is biased negatively if $(1/t)\sigma^2_{abc}$ is less than the variance due to the interaction of that effect with the effect which is common to both double interactions. For example, if AB, AC, and ABC are present, negative F test bias will occur in the test of A if $(1/t)\sigma^2_{abc}$ is less than $\sigma^2_{ac} + \sigma^2_{ab}$; in the test of B if $(1/t)\sigma^2_{abc}$ is less than σ^2_{ab} ; and in the test

of C if $(1/t)\sigma^2_{abc}$ is less than σ^2_{ac}. Of course, positive bias results if the inequality is reversed.

Case 7. All Interactions

All tests are negatively biased unless the sum of the variances for the three double interactions is less than the variance due to the triple interaction.

In summary, in the zero random variate model all tests except one are biased if one or more interactions appear. The unbiased result occurs when one double interaction is extant and the test is on the variate not included in the interaction. However, because of the presence of the fixed effect interaction in both the numerator and denominator, this unbiased test is not a valid F test.

One Random Variate Model

The paradigm is shown in Table 2. With the introduction of a random effect (C) more components appear, with σ^2_{ac} and σ^2_{bc} included as a component for A and B, respectively. Furthermore, the coefficients of σ^2_{abc} change to $t - (2/t)$ for the residual and to $t - (1/t)$ for the random effect. The coefficients remain as $t - (2/t)$ for the two fixed effects.

TABLE 2

One Random Variate Model

Variates	EMS (Coefficients of)							
	σ_e^2	σ_{abc}^2	σ_{bc}^2	σ_{ac}^2	σ_{ab}^2	σ_c^2	σ_b^2	σ_a^2
A	1	(t-2)/t	1	1	0	0	0	t
B	1	(t-2)/t	1	1	0	0	t	
C*	1	(t-1)/t	0	0	1	t		
Residual	1	(t-2)/t	1	1	1			

* Variates with asterisk in Tables 2-4 are random main effects.

Case 1. One Double Interaction

If the interaction involves the fixed effects, negative F test bias occurs in tests of these two but no bias occurs in the test of the random effect. However, this latter is not distributed as F. If the interaction involves the random effect, only the test of the random effect is negatively biased. However, of the

two unbiased tests only that of the variate interacting with the random effect is a valid F test.

Case 2. *Triple Interaction*

The F tests of the two fixed effects are free from bias, but are not distributed as F. In the test of the random effect positive bias occurs.

Case 3. *Two Double Interactions*

If one interaction involves a random effect and the other contains both fixed effects, all F tests are negatively biased. If both interactions involve the random effect, negative bias occurs in the test of the random effect. The test of the two fixed effects are unbiased but are not distributed as F. Again we have a ratio of two noncentral chi square statistics whose distribution depends on the parameters of the fixed effects interaction included in the mean square.

Case 4. *Three Double Interactions*

All tests are negatively biased, but the bias is greater for the test of the random effect than for the tests of the two fixed effects because the random effect contains one interaction term while each of the fixed effects contains two.

Case 5. *One Double Interaction and the Triple Interaction*

If the fixed effects interact, then positive bias occurs in the test of the random effect but negative bias appears in the tests of the fixed effects. If the double interaction includes the random effect, then the tests of the two fixed effects are unbiased but not valid F tests. The test of the random effect is biased, with negative bias if σ^2_{ac} or σ^2_{bc} is greater than $(1/t)\sigma^2_{abc}$, or with positive bias if the reverse occurs.

Case 6. *Two Double Interactions and the Triple Interaction*

If one of the double interactions involves the random effect and the other does not, the tests of the fixed effects are negatively biased; the test of the random effect is negatively biased if the variance of the interaction including the random effect is greater than $(1/t)\sigma^2_{abc}$, otherwise positive bias occurs. If both double interactions include the random effect, the F tests for the fixed effects are unbiased but are not valid F tests. The test of the random effect is negatively biased if $\sigma^2_{ac} + \sigma^2_{bc}$ is greater than $(1/t)\sigma^2_{abc}$; otherwise positive bias occurs.

Case 7. *All Interactions*

Tests of the fixed effects are negatively biased; the test of the random effect is negatively biased if $\sigma^2_{ac} + \sigma^2_{bc}$ is greater than $(1/t)\sigma^2_{abc}$; otherwise positive bias occurs.

In summary, with one random effect included in the experiment more tests are unbiased even though interactions occur and some tests are distributed as F.

Two Random Variates Model

The paradigm for this model is shown in Table 3. With the appearance of another random effect (B) the coefficients of σ^2_{abc} change to $t - (1/t)$ for all four effects, and σ^2_{ab} is included for A and σ^2_{bc} for C.

TABLE 3

Two Random Variates Model

Variates	EMS (Coefficients of)							
	σ_e^2	σ_{abc}^2	σ_{bc}^2	σ_{ac}^2	σ_{ab}^2	σ_c^2	σ_b^2	σ_a^2
A	1	(t-1)/t	1	1	1	0	0	t
B*	1	(t-1)/t	1	1	0	0	t	
C*	1	(t-1)/t	1	0	1	t		
Residual	1	(t-1)/t	1	1	1			

Case 1. *One Double Interaction*

If the interaction involves the two random effects, all tests are unbiased but only the tests of the random effects are distributed as F. If the interaction includes the fixed effect, two tests are unbiased, the test of the fixed effect and that of the random effect not included in the interaction. The test of the fixed effect is a valid F test but the F ratio for testing the random effect is not distributed as F. The random variate not included in the interaction is negatively biased.

Case 2. *Triple Interaction*

All F tests are unbiased but only the test of the fixed effect is valid and distributed as F.

Case 3. *Two Double Interactions*

If the interactions that are present both involve the fixed effect, tests of the random effects are negatively biased. The test of the fixed effect is unbiased and a valid F test. If one of the interactions includes both random

effects, the random effect included in both interactions is negatively biased but the tests of the other two effects are unbiased but not distributed as F.

Case 4. *Three Double Interactions*

The tests of both random effects have negative bias; the test of the fixed effect is unbiased but not a valid F test.

Case 5. *One Double Interaction and the Triple Interaction*

If the double interaction involves both random effects, all tests are unbiased but not distributed as F. If the double interaction contains the fixed effect, the test of the random effect involved in the interaction is negatively biased; the test of the two other effects are unbiased but only the test of the fixed effect is distributed as F.

Case 6. *Two Double Interactions and the Triple Interaction*

If the fixed effect is involved in both double interactions, the test of the fixed effect is a valid F test but the tests of the remaining effects are negatively biased. If only one interaction includes the fixed effect, the random effect that is included in both double interactions is negatively biased and the tests of the other effects are unbiased but are not valid F tests.

Case 7. *All Interactions*

Tests of the random effects are negatively biased; the test of the fixed effect is unbiased but not a valid F test, because a component (σ^2_{bc}) which does not involve the fixed effect is included in the mean square of the fixed effect.

In summary, if two random effects are included in the experiment one or more tests are unbiased; the test of the fixed effect is usually unbiased and distributed as F. All biased tests are negative.

Three Random Variates Model

The paradigm is shown in Table 4. With all random effects present the coefficient for σ^2_{abc} is unity for all effects. All interaction components appear in the mean square of the three main effects as well as in the residual, indicating that all tests will be unbiased, but not necessarily valid.

Case 1. *Double Interaction*

Only the tests of the effects included in the interaction are valid.

Case 2. *Triple Interaction*

All tests are valid.

Case 3. *Two Double Interactions*

Only the effect included in both interactions provides a valid F test.

TABLE 4

Three Random Variates Model

Variates	EMS (Coefficients of)							
	σ_e^2	σ_{abc}^2	σ_{bc}^2	σ_{ac}^2	σ_{ab}^2	σ_c^2	σ_b^2	σ_a^2
A*	1	1	1	1	1	0	0	t
B*	1	1	1	1	1	0	t	
C*	1	1	1	1	1	t		
Residual	1	1	1	1	1			

Case 4. *All Double Interactions*

No tests are distributed as *F*.

Case 5. *One Double Interaction and the Triple Interaction*

The results are the same as in Case 1, i.e., only the tests of the effects included in the double interaction are valid.

Case 6. *Two Double Interactions and the Triple Interaction*

Only the test of the effect which is included in all interactions is a valid *F* test.

Case 7. *All Interactions*

No tests are valid.

Thus when all effects included in the experiment are random variates all tests are free from bias regardless of how many interactions are present. Furthermore, more valid tests occur than in the previous models.

In conclusion, it is interesting to note that as the number of random variates increases the number of unbiased *F* tests increases likewise until in the three random variates model all tests are free of bias. Paralleling this trend is an increasing number of valid *F* tests as well. In the first two models negative bias occurs most frequently but small positive bias is possible when the triple interaction is present. In the third model all biases are negative. Thus as the number of random variates increases, more tests are insensitive to deviation from the traditional assumptions that interactions must be zero. The most frequent occurrence in psychological research with the Latin square design probably is represented by the one or two random variates model in which subjects represents one of the effects. Contrary to previous

thought, less bias will occur with interactions present inasmuch as some tests will be unbiased; however, a fewer number will be valid. If one were to use the derivation technique of Mood and McNemar, more tests would be unbiased and valid in the one and two random variates models since more components are included in the expected value of mean square of the three main effects by this procedure.

Furthermore, contrary to McNemar's assertions, and in agreement with Gourlay, most of the tests have negative bias. This indicates that the probability of a Type II error increases with these Latin square designs for some tests of significance. Therefore, if the investigator reports a significant result in most of the tests within the first two models, or in any of the tests of the third model, he can be safe in stating that the probability of such an event is at or below the probability level chosen. However, if significance is not indicated by some tests when an interaction is present (and unknown to the investigator), the stated results are not as certain. It appears that prior to using a Latin square design the investigator should familiarize himself with the various cases and be aware of possible distortions which may occur.

REFERENCES

[1] Anderson R. L. and Bancroft, T. A. *Statistical theory in research.* New York: McGraw-Hill, 1952.
[2] Gourlay, N. *F*-test bias for experimental designs of the Latin square type. *Psychometrika*, 1955, **20**, 273–287.
[3] Greenwood, J. A. Analysis of variance and components of variance factorial experiments. Unpublished paper, Bureau of Aeronautics, 1956 (revised).
[4] Kempthorne, O. *The design and analysis of experiments.* New York: Wiley, 1952.
[5] Lindquist, E. F. *Design and analysis of experiments in psychology and education.* New York: Houghton Mifflin, 1953.
[6] McNemar, Q. On the use of Latin squares in psychology. *Psychol. Bull.*, 1951, **48**, 398–401.
[7] McNemar, Q. *Psychological statistics.* New York: Wiley, 1955.
[8] Mood, A. M. *Introduction to the theory of statistics.* New York: McGraw-Hill, 1950.
[9] Tukey, J. W. Interaction in a row by column design. Memorandum Report 18, Princeton Univ., 1949.
[10] Wilk, M. B. and Kempthorne, O. Non-additivities in a Latin square design. *J. Amer. statist. Ass.*, 1957, **52**, 218–236.

Manuscript received 10/9/57
Revised manuscript received 2/24/58

PSYCHOMETRIKA—VOL. 23, NO. 4
DECEMBER, 1958

A MODIFICATION OF KENDALL'S TAU FOR MEASURING ASSOCIATION IN CONTINGENCY TABLES

BERTRAM P. KARON AND IRVING E. ALEXANDER
PRINCETON UNIVERSITY

A coefficient of association τ' is described for a contingency table containing data classified into two sets of ordered categories. Within each of the two sets the number of categories or the number of cases in each category need not be the same. $\tau' = +1$ for perfect positive association and has an expectation of 0 for chance association. In many cases τ' also has -1 as a lower limit. The limitations of Kendall's τ_a and τ_b and Stuart's τ_c are discussed, as is the identity of these coefficients to τ' under certain conditions. Computational procedure for τ' is given.

Consider a contingency table consisting of two sets of ordered categories in which the numbers of categories within each set or the numbers of cases within each category are not identical. Such sets of ordered categories may be thought of as rankings with ties. It is well known that Kendall's τ [1] may be applied to contingency tables treating the categories as tied ranks. However, limitations in its application have recently been pointed out [3].

In the non-tied case, Kendall's coefficient has the following desirable properties.

(i) If agreement between two rankings is perfect, $\tau = +1$.
(ii) If inverse agreement between two rankings is perfect, $\tau = -1$.
(iii) If there is chance agreement between the two rankings $E(\tau) = 0$.
(iv) The sample τ is an unbiased estimate of the parametric value for the population from which the sample is randomly drawn.
(v) The sampling properties of τ are known and statistical test procedures are available.

One way of defining τ for the case where there are no ties is

$$\tau = \frac{S}{S_{\max}}. \tag{1}$$

S is the number of pairs of objects which are in the same order in both rankings minus the number of pairs of objects in which the order is reversed; $S_{\max}$ is the maximum value attainable by S when agreement is perfect.

Since for n objects there are $n(n-1)/2$ pairs of objects the maximum attainable value of S is $n(n-1)/2$. Thus τ may be written

$$\tau = \frac{S}{\frac{1}{2}(n)(n-1)}, \tag{2}$$

which clearly varies from $+1$ to -1.

If there are one or more ties in either ranking, (1) is no longer identical with (2), since pairs which are tied are not counted in S. Tied objects are considered to have no order with respect to each other and therefore their order in the ranking in which they are tied can neither agree nor disagree with their order in the other ranking. Thus (2) can never reach $+1$ or -1. For such cases, Kendall defines (2) as τ_a, where suggested usage is confined to the situation in which the ties arise from an inability of the ranker to comply with instructions to carry out a complete ranking.

Kendall then defines τ_b to account for all other cases of tied ranks,

$$\tau_b = \frac{S}{[\frac{1}{2}n(n-1) - T]^{1/2}[\frac{1}{2}n(n-1) - U]^{1/2}}. \tag{3}$$

T may be written

$$T = \tfrac{1}{2} \sum_r t_r(t_r - 1), \tag{4}$$

where t_r is the number of objects tied for rank r of one of the rankings, and

$$U = \tfrac{1}{2} \sum_s u_s(u_s - 1), \tag{5}$$

where u_s is the number of objects tied for rank s of the other ranking. For any rank q in which there are no ties,

$$t_q(t_q - 1) = 0 = u_q(u_q - 1). \tag{6}$$

Therefore the summations in (4) and (5) need be taken only over the tied ranks, and when no ties occur in either ranking, (3) reduces to (2).

If the number of categories in each ranking and the number of cases in each category are the same in both rankings, τ_b can attain $+1$; if the number of categories is the same and the number of cases in each category is the same in reverse order, τ_b can attain -1; if the number of categories is the same, and there is symmetry such that the number of cases in each category is the same in either order, τ_b can attain both limits.

Stuart [3] notes that even τ_b cannot attain in all cases the limits of $+1$ and -1. Stuart defines a coefficient τ_c for the general $r \times s$ contingency table,

$$\tau_c = \frac{S}{n^2(m-1)/(2m)}, \tag{7}$$

where m is the number of rows or the number of columns, whichever is smaller. If n is an even multiple of m, τ_c can attain the limits $+1$ or -1 when all the cases are confined to a longest diagonal, and the number of cases in each cell of that diagonal are equal. If n is not an even multiple of m then τ_c cannot reach these limits, but, according to Stuart, τ_c will approach them closely for large n.

However, for small n (when n is not a multiple of m) the discrepancy

is considerable, and τ_c is not satisfactory. More serious is the fact that even when n is a multiple of m, and even when n is large, τ_c cannot attain either $+1$ or -1 except for the special case when the number of objects in each of the categories of the two rankings are arranged to allow the number of cases in each diagonal cell to be equal. Such special cases are not common.

A simple coefficient for all sets of ordered categories, one that includes the desired properties, is readily available. Equation (1) may be used to define a coefficient τ'. It is evident that for all situations, τ' can reach a maximum of $+1$ only for as perfect an agreement as is attainable given the particular $r \times s$ table, and has an expected value of 0 when there is a chance agreement. It is also clear that τ_a , τ_b , and τ_c are all equal to τ' in the cases where they have the desired properties outlined for τ. In cases where no formula is readily available for $S_{\max}$, it is simple to calculate numerically. All that is necessary is to fill in the $r \times s$ table under consideration as the observations would occur if association were perfect, and compute S in the same fashion as one computes the observed S.

To illustrate the computation of τ' consider an example from Macht [2]. From the theory to be tested, 10 objects are ranked. The criterion is such that only the first 5 may be ranked, while the other 5 objects are placed into one ordered category.

		Theory									
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Criterion	*1*	0	0	1	0	0	0	0	0	0	0
	2	1	0	0	0	0	0	0	0	0	0
	3	0	1	0	0	0	0	0	0	0	0
	4	0	0	0	1	0	0	0	0	0	0
	5	0	0	0	0	0	1	0	0	0	0
	6	0	0	0	0	1	0	1	1	1	1

In this sample the rows represent categories ordered according to the criterion and the columns represent the ranks according to the theory. Therefore, if an observation falls in a row below that of another observation it is ranked lower by the criterion, and if an observation falls in a column to the right of that of another observation, it is ranked lower by the theory. If an observation is both below and to the right of another, it is ranked lower by both the criterion and the theory, and the comparison between the two is in agreement on both rankings. If an observation falls below, but to the left of another, it is ranked lower by the criterion but higher by the theory, and the comparison between the two is in disagreement. If an observation falls in the same row or in the same column as another, it is tied on either the criterion or the theory, and therefore the comparison between the two can neither agree nor disagree. In order not to make the same comparison

more than once, each observation may be compared only with those which fall in rows lower than its own. Thus, one would compare the observations in the first row with those in all other rows. When the observations in the second row are considered, they would be compared only with observations in the third row, fourth row, fifth row, etc., but they would not be compared with observations in the first row, since these comparisons would already have been made.

Now S may be computed, as indicated by Kendall, in the following manner.

(i) For each cell of the table, count and record the number of observations which fall in cells which are both below and to the right of that cell.
(ii) Subtract the number of observations which fall in cells both below and to the left of that cell.
(iii) Do not count or subtract any observations which lie directly below, or fall on the same horizontal line, or lie above the cell.
(iv) Multiply the result for each cell by the number of cases in that cell.
(v) Sum the results for each cell over the whole table.

For this example,

$$S = (7 - 2) + 8 + 7 + 6 + (4 - 1) = 29. \tag{8}$$

To determine $S_{\max}$, write a contingency table showing perfect prediction of the criterion by the theory.

		Theory									
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Criterion	*1*	1	0	0	0	0	0	0	0	0	0
	2	0	1	0	0	0	0	0	0	0	0
	3	0	0	1	0	0	0	0	0	0	0
	4	0	0	0	1	0	0	0	0	0	0
	5	0	0	0	0	1	0	0	0	0	0
	6	0	0	0	0	0	1	1	1	1	1

Here,

$$S_{\max} = 9 + 8 + 7 + 6 + 5 = 35. \tag{9}$$

τ' can vary from $+1$ to -1, and for this example

$$\tau' = \frac{S}{S_{\max}} = \frac{29}{35} = .83. \tag{10}$$

It is interesting to note the values reached by τ_a, τ_b, τ_c, and τ' when the theory predicts the criterion perfectly, and when consequently a satisfactory coefficient should equal 1.

(11) $\tau_a = .78;$

(12) $\tau_b = .88;$

(13) $\tau_c = .84.$

(14) $\tau' = 1.00.$

τ' seems the most satisfactory coefficient. It always has $+1$ as an upper limit indicating perfect association, and it always has 0 as its expected value when there is a chance association, since its numerator S has an expected value of 0. In addition, it will have -1 as a lower limit indicating perfect negative association for those situations in which $S_{\min} = -S_{\max}$. Where $S_{\min} \neq -S_{\max}$, if an attainable lower limit of -1 for perfect negative association is more important than an attainable upper limit of $+1$ for perfect positive association, $-S_{\min}$ may be used in the denominator of (1) in place of $S_{\max}$.

It should also be noted that a test of significance for the hypothesis of no association exists, since, for all tau statistics, the test of significance is based on the distribution of S and not of τ. Thus the change in the denominator in defining τ' affects only the measurement of association, and not the test of significance. The distribution of S under the hypothesis of no association has been discussed by Kendall [1].

REFERENCES

[1] Kendall, M. G. *Rank correlation methods.* London: Griffin, 1948.

[2] Macht, L. B. An application of the resultant weighted valence theory of level of aspiration to the study of occupational preference. Unpublished undergraduate thesis, Princeton Univ., 1957.

[3] Stuart, A. The estimation and comparison of strength of association in contingency tables. *Biometrika,* 1953, **40**, 105–110.

Manuscript received 10/15/57
Revised manuscript received 1/28/58

BOOK REVIEWS

GEORGE S. WELSH AND W. GRANT DAHLSTROM. *Basic Readings on the MMPI in Psychology and Medicine.* Minneapolis: University of Minnesota Press, 1956. Pp. xvii + 656.

A collection of 66 papers and a 698-item bibliography provide a systematic compilation of representative information on the most prominent of all personality inventories. The collection is an excellent representation of papers interesting to the clinician, covering the basic validations of the instrument, the principal papers on fringe scales for dominance, ego-strength, and so on, and naturalistic reports on patient groups whose diagnoses range from alcoholism to cancer. The articles are skillfully edited to avoid duplication, and new papers written specially for the volume fill critical gaps.

The reports of the scale development present a fascinating example of the research process in the hands of flexible investigators who can be honest with themselves. The first papers (ca. 1940) optimistically embarked on developing quantitative discriminant scales for psychiatric diagnosis. As the validities proved disappointing, trial and error was used in the hope of improvement. One hopeful attack introduced suppressor items, and later (1946) the *K* scale to correct for test-taking attitudes. Even the corrected scales were not very much in agreement with diagnoses, and from this date forward the papers increasingly deny that prediction of such criteria is or should be the function of the test. The most meaningful subsequent research is directed to connecting scales and patterns with general descriptive constructs.

Considering the prominence the *MMPI* attained by virtue of its timely appearance, just as the war thrust new demands upon clinical psychology, it is incredible that its foundations are so shaky. The authors, though painstaking, established weights for items on tiny clinical samples. The sample *N*'s for original item selection were as follows: *HS*, 50; *D*, 50; *PT*, 20; *Hy*, several samples; *Ma*, 24; *PD*, 100 plus a second sample of unstated *N*; etc. The fact that *MMPI* scales have worked at all on cross-validation conflicts sharply with the opinion of many authorities that samples for establishing item weights should be 500 or larger. The authors freely criticized their own scales, repeatedly using such words as "disappointing" and "weak." Three of the main scales, indeed, were released with the intention of replacing them later.

Hindsight indicates that the inventory was designed quite inefficiently. The *T-F-?* pattern invites maximal distortion by response sets; the indiscriminate mixing of obvious and subtle items prevents the interpreter from capitalizing on the virtues of either type; the weighting of items which differentiate patient groups from normals may leave wholly out of consideration items which differentiate patient groups from each other, and leads to undesirably high scale intercorrelations. [In a patient sample, eight intercorrelations out of 36 are between .60 and .86 (p. 259)]. Moreover, the generally very low intercorrelations between *MMPI* scales and the *CPI* scales derived by Gough from essentially the same items indicate that the basic *MMPI* scores extract only a small fraction of the information in the responses.

The papers presented show consistent but weak relations between *MMPI* and behavior. The research reports are too little integrated to indicate just how much confidence can be placed in any particular interpretation. A number of studies suggest that clinicians assign patients to categories with about 70 per cent success, in experiments where chance success would be 50 per cent. None of the studies in this volume bear on the more pertinent question of success when realistic base rates are taken into account.

This volume is not a collection of model studies. Few studies contain gross errors, but many of them reflect designs and conceptualizations which have become outmoded in the last decade. The investigations which seem most meritorious in terms of their sound-

ness and informativeness are those by Black on college girls (p. 151), Peterson on predicting hospitalization of outpatients (p. 407), Schiele and Brozek on experimental starvation (p. 461), and Barron on ego-strength (p. 579). Two of the four are written for this volume. Some older studies show several faults in addition to neglect of base rates and use of small samples. In studies of group differences the significance tests outnumber the subjects as much as 20 to 1; under these conditions, probability values are meaningless. "Signs" are generally cross-validated in a suitable manner, but in one instance (p. 311) the signs are revised on the second sample and significance is then tested on the same cases for the revised signs. Elsewhere (p. 330) signs are claimed to give a significant χ^2 for selection of teachers, but correct application of a 2 x 2 table would show nonsignificance. Outside of two papers by Gough, insufficient recognition is given to facade or "hello-goodbye" effects.

In this book, as in other *MMPI* literature, great emphasis is placed on "patterns." This term is applied indiscriminately, and the experimental designs rarely bear on the conclusions drawn. Conclusions of the form "*Pt* is higher than *Sc* in group X" should be tested by showing what proportion of group X has $(Pt - Sc) > 0$. The majority of studies make such interpretations from inspection of the mean difference, and compute significance by showing that one or both scores taken separately differ from the normal—a finding irrelevant to the conclusion. As in this example the so-called "patterns" are often no more than simple difference scores. The writer would advocate reserving the terms "pattern" and "configuration" for conjunctive, nonlinear formulas. Where nonlinear formulas are offered (for example in Welsh's internalization ratio), the claim for configural validity ought to be supported by showing that this statistically awkward form is actually more valid than a suitable linear composite. The study by Little and Shneidman (p. 332), the most truly configural in the book, requires special criticism. A *Q*-sort of a single case was made by *MMPI* judges and *Q*-correlated with the sort made on the basis of the clinical case record. The average validity is said to be .67. Such an index is meaningless by itself, since selection of statements could swing the validity in either direction by almost any amount. In this case it reflects chiefly an impression of bad adjustment (four *T*-scores are over 90!) rather than a "personality pattern." The *Q*-sort for *any other patient* with manifest extreme disturbance would surely correlate highly with the criterion for this man. As a minimum, the authors should show how the correlation for *A*'s *MMPI* sort with *A*'s criterion sort compares to correlations of *MMPI* sorts for patients *B*, *C*, and *D* with *A*'s criterion sort.

One can only commend the compilers for providing so adequate a picture of *MMPI* history. A reader is left with respect for the progress we have made, particularly in discarding over-optimistic expectations. He should also be dismayed that so much conscientious effort on our most carefully planned personality inventory leaves us in our present state. Interpretation of *MMPI* profiles rests far more on the interpreter's "experience" than on validated principles. Within any typical intake group in a clinic or a student body, it seems quite doubtful that one can consistently make dependable inferences about an individual's degree of disturbance or personality structure. It is fortunate that so many of these papers urge that the test be used only tentatively, as a supplement to other means of investigating the individual.

University of Illinois LEE J. CRONBACH

PHILIP H. DUBOIS. *Multivariate Correlational Analysis.* New York: Harper & Brothers, 1957. Pp. xv + 202.

This book is concerned with the descriptive statistics of multiple linear regression systems. It appears to have been written chiefly for the practicing research worker in the

social sciences who has a sound grasp of elementary descriptive statistics. Most of the topics are presented from a unifying viewpoint most clearly expressed in what the author calls the basic theorem—actually a lemma—of correlational analysis:

> Every value of a variate may be divided uniquely into two uncorrelated components: a portion perfectly correlated with an outside variate and a portion uncorrelated with the outside variate.

Not only does this lemma serve to unify the discussion of various topics and procedures, but also to guide the research worker in their use.

Valuable as the lemma is, as a unifying conception, the major focus of the book is upon the regression *techniques*, which explicitly exhibit the lemma in action. This is achieved by the author's development of Yule's partial covariance formulation into a general computing routine, based on the method of single division.

In the first nine sections, multiple, partial, and part correlation are discussed in terms of the partial covariance concept and computing routine. With the appropriate adaptation of the routine, forward solutions yield directly multiple R, partial correlations of any order, multiple-partial, and multiple part correlations. Corresponding regression coefficients may be found by completing the back solution. The general computing is simple, efficient, and includes systematic computational checking procedures. The author also presents an abbreviated method, and a Wherry-Doolittle type of procedure for selecing a subset of predictors.

The procedures consist of systematic elimination of variance associated with each predictor. At each stage of the analysis partial variances and covariances for remaining "residual" variables are exhibited. This conception of correlational analysis is particularly advantageous in understanding correlations involving residuals, such as partial, part, multiple-partial, and multiple part correlations. This is the clearest presentation of these topics seen by the reviewer. With the exception of partial correlation, it is not uncommon to see these matters dismissed in a brief paragraph containing a formula with no discussion of meaning or use; the reader is left to interpret juggled subscripts. The author deals with these subjects effectively and insightfully.

Correlations involving residuals may be used to introduce controls statistically in situations where they cannot be introduced experimentally. As an example of the use of part correlation, the author cites the frequently encountered problem where individual differences on the criterion variable existing prior to training obscure the relation between an aptitude predictor and post-training performance. The problem is to remove the effect of initial level of performance from the measurement of final level, but not from the aptitude predictor. The same logic applies to the use of the multiple part correlation for the case of more than one predictor. Similar procedures may be used to purify a criterion of variance irrelevant to the purpose of an investigation.

Two sections are devoted to factor analysis. No attempt is made to present a systematic discussion of this topic, but rather relations are demonstrated between factor analysis and multiple regression systems, as analyzed in terms of partial covariances. A method is presented in which pivoting is done on a component defined by variance common to at least two variables. Starting with uni-factor reference variables for the first factor, two-factor reference variables for the second factor, etc., the analysis proceeds in terms of partial covariances by an adaptation of the general computing routine. Procedures are given for picking factor-defining variables at each stage of the analysis.

This method of factor analysis leads to a definition of factors which can be interpreted invariantly across equivalent samples if suitable reference variables are introduced in each analysis. The procedures lead to a definition of the common factor space as a by-product of the analysis, thus circumventing the problem of initial communality estimation. Since factors are defined serially in terms of residual covariances, extracted factors are orthogonal. Futher extraction and rotation may be introduced, if desired. Although uni-

versal applicability of the method is not claimed, it is a useful one when dealing with rather well understood domains in applied psychology. The computational effort is judged to be approximately equivalent to that of the centroid method.

Readers with primary interest in statistics may prefer to start with a section entitled "Some Mathematical Considerations." This section develops and summarizes in mathematical terms relations discussed in the first ten sections. However, it is not at all formidable for the less sophisticated reader. The section closes with a discussion of the relationship between the author's and Yule's formulas. A number of topics receive very brief treatment in a section on inference and prediction. Included are corrections for attenuation, inferring by factorial extension correlations missing from a matrix, alienation and the standard error of estimate, shrinkage, the reliability of a residual variable, and the standard error of multiple correlations. A final section deals with the role of correlation analysis in social science research. The role is illustrated with six uses of the basic lemma. There follows a brief discussion of closed systems such as those generated by forced choice scores.

In spite of its title, the scope of this book is limited by exclusion of many topics falling in the general domain of correlational analysis. One will not find any discussion of pattern analysis, multivariate models for norming and scaling, corrections of correlation matrices for effects of selection on more than one variable, canonical regression, image and radex models, discriminant functions, or analysis of dispersion. Within the scope chosen by the author, concepts and working procedures are clearly presented. Some readers may prefer some geometric presentation to supplement the verbal and algebraic approach used. The book can be used effectively as a text in courses with similar scope and level; however, many instructors will want to supplement the text with more examples of situations requiring the use of the procedures and with more thorough discussion of some of the topics presented.

The format of the book includes a list of 36 references, a table of R as a function of partial variance, that portion of a square root table most useful in the general computing routine, and a computing chart with directions for its use. A useful glossary and name and subject indices complete the format. The publisher's use of linotype Caledonia makes an attractive and readable page. Typographical errors are rare and of minor consequence. In summary, this book, though limited in scope, will be useful and clarifying for many readers. Regular readers of this journal will probably find much of the discussion elementary, but may nevertheless profit from looking at familiar things from a refreshing viewpoint.

Wright Air Development Center JOHN A. CREAGER
Lackland Air Force Base, Texas

JOHN B. MINER. *Intelligence in the United States.* New York: Springer Publishing Company, 1957. Pp. xii + 180.

About a third of American students below the tenth grade and over nine years old could be doing work at college sophomore level or above. Such an assertion is arresting in a period of concern over utilization of intellectual resources. The contributions offered in this survey are, first, information on the distribution of verbal ability in the American population drawn from a cross-sectional national sample. Further, these novel data are compared with distributions in a model society with a perfect correlation between verbal ability and educational level or occupation. A statistical analysis of the reshuffling required in different subgroups yields information on the location of talent reserves.

Why verbal ability? The study is a somewhat accidental outgrowth of standardization research requiring a fifteen-minute doorstep intelligence test. Not unwisely, the author chose a vocabulary test as a close approximation in the time available. It would

be hard to dispute the author's argument that verbal ability is more crucial to success in our educational hierarchy than is any other special ability. His use of the term *intelligence test* to describe the measure is objectionable if the book has a nonprofessional audience. Probably a vocabulary test is even more sensitive to educational effects than most general intelligence tests. Although Miner warns that environmental stimulus potential is crucial to the level of functioning at a given point, his own perference for the term intelligence test suggests an underplaying of this factor.

This is the first cross-sectional national sample of ability available. It is unfortunate that the sampling technique used—a combination of cluster and quota sampling—will not permit establishing confidence limits. The original sample of 1500 deviated sufficiently from census characteristics to require additional quota-selected subjects to the number of a quarter of the final sample, suggesting that there was some bias in sample selection. As a result there are unknown effects on cell distributions even when marginal frequencies have been matched with the census, and unknown effects on the ability measure. Since more active, younger, more talkative respondents usually appear in quota samples one would guess that the estimates of ability are biased upward. There may be no differential effects on subgroups in the samples, however, and for Miner's analysis differential effects are crucial.

Even had the sample been drawn with appropriate known-probability methods, the standard error would be systematically underestimated, since corrections for clustering were not used. In vitiation of these shortcomings one can note that the differences Miner found are generally substantiated by other studies of selected samples, such as the Army recruit studies and Wolfle's school AGCT samples.

Differences appeared in relation to education, occupation, rural-urban residence, geographical area, race, class identification, and religion. All these differences are compared to those found in other studies with coverage of relevant literature. One of the useful findings that appears repeatedly throughout the analyses is that the verbal knowledge distribution in housewives and retired persons does not deviate significantly from that of persons in the labor force, when educational differences are controlled in the older persons.

Miner next proceeds to an analysis of discrepancies from the perfect correlation model, using two methods. The tenth percentile of vocabulary scores is taken as the minimum current entry level for any given school grade or occupational stratum, with occupations divided into four skill strata. Those currently placed below that level who have scores above the minimum entry point for a higher level are considered underplaced. Since the correlation of verbal knowledge with both education and occupation is low, there is inevitably a large proportion in the lower strata who are underplaced. But Miner is not interested in over-all underplacement so much as in differential underplacement in various subgroups.

Thus, whites are found to be more often educationally underplaced than negroes, and unskilled underplaced more than highly skilled workers. Within the occupational system, the less educated and working class persons could more often take a more highly skilled occupation in stride.

Since cross-tabulations are not supplied, it is not altogether clear why these discrepancies exist. A group would have less underplacement if (*a*) there is a higher correlation with vocabulary than in the total sample; (*b*) the group is highly placed already; (*c*) the group is more homogeneous than the total sample and is low on both measures. Looking at the vocabulary data it appears that (*c*) applies to negroes, and (*b*) to highly skilled persons.

The second technique for analysis of efficiency in talent distribution consists of postulating a system in which there is a perfect correlation between verbal ability and educational or occupational stratification, with no changes in the numbers at each level.

On the conservative assumption of no changes in skill distribution, there would be demotions as well as promotions to achieve the ideal system. Given these assumptions, the cut-off points are now raised because of the higher standards in a nonoverlapping stratification system. Since promotion is now available only to those at the top of their present level, a shift in the subgroup difference results occurs.

Naturally, educational promotion rates are greater for the lower grades and demotion greater at higher levels because of ceiling effects. For instance, 3.9 per cent of the students below ninth grade would be promoted to college sophomore level or higher, and 21.7 per cent of the college students would be demoted in compensation. Differential results appear only for sex, with more girls being demoted, corroborating Wolfle's finding of relatively higher grades for girls with AGCT controlled.

In the occupational system, women would be more often promoted and men demoted, possibly due to underemployment of women in the highest skill level. More demotions would occur in the higher education and class levels than in the lower groups.

The one apparently paradoxical result in these analyses is the finding that negroes are not relatively more underplaced, but on the contrary are in some cases less underplaced than whites. Here is one of the pitfalls of Miner's method, which he notes but cannot surmount. He is comparing two measures both influenced by environmental stimulus potential and is trying to interpret discrepancies. As long as the chief contributor to variance in the verbal knowledge measure is individual ability, discrepancies can be sensibly interpreted and policy implications considered. Such is probably not the case with negro scores. In the educational analysis, the norms for a given grade are derived from pooling all schools. It is quite probable that standards for schools which negroes attend are sufficiently beneath the white averages to lower vocabulary exposure for negroes and thus reduce underplacement proportions. Anastasi's New York studies suggest that additional discrepancies in verbal exposure histories predate school experience.

What does it mean then to say that adult negroes are by and large not capable of further education? The core of the contradiction lies in the contribution of education itself to the measure of capacity for education, and the failure to equate successfully past education. The same difficulty appears in analysis of adult educational potential when it is assumed that to be capable of a college education an adult with a grade school education should have the same verbal ability scores as students now in college. Miner emphasizes motivation and education in his treatment of group differences in verbal knowledge; he notes both in the introduction and summary that he is talking about current and not in any sense potential functioning. But his underplacement results are often phrased in a manner that obscures this qualification, which, from a public policy standpoint, is of great importance.

Miner's data is of more interest than the policy suggestions which he proposes without any extensive considerations of alternative factors (which do, and in some cases should, reduce the correlation of verbal ability and education or occupation). In his educational utopia, grade placement would entirely ignore chronological age and be based on achievement criteria. The problems raised in such a system are dismissed rather summarily although this kind of proposal has been controversial for some time. Of perhaps greater interest from the view of public policy is the evidence by one criterion of considerable underusage educationally and occupationally of large numbers of persons both in the labor force and in retirement. In Miner's society there is no reward for effort, no pay-off from early independence training and high need achievement, and the Betas do not aspire beyond their abilities. He presents the hope that guidance counselors and company personnel men can selectively counteract these influences in the real world.

SUSAN M. ERVIN

University of California, Berkeley

ROBERT R. BUSH, ROBERT P. ABELSON, AND RAY HYMAN, *Mathematics for Psychologists, Examples and Problems*. New York: Social Science Research Council, 1956. Pp. iv + 86.

This paperbound volume, prepared for the Social Science Research Council during the summer of 1954 under Bush's direction, provides a fund of examples and problems illustrating mathematical applications in psychology. These are chosen and classified so that they can be used in four of the standard undergraduate mathematics courses. The book is not, and was not intended to be, either a systematic treatise on mathematics or on the uses of mathematics in psychology; rather, it presents specific illustrations, drawn from the psychological literature, of applications of some of the more familiar mathematical topics. No really elaborate developments are presented *in toto*. Although its main use undoubtedly will be to supplement mathematics texts, it should also aid those preparing courses on mathematical psychology.

Within their chosen framework, the authors have done an effective job. The coverage, although not intended to be exhaustive, is broad, the references to the literature are generous (124 items in the bibliography), and the writing is concise and clear. My only question is whether teachers of mathematics will not find the descriptions of the underlying psychological problems too abbreviated. Probably they will be forced to read some of the research literature before they will feel reasonably confident in employing these examples; quite possibly this will serve a desirable long range purpose, and certainly with this volume in hand teachers of mathematics will know where to read.

Each of the four main sections of the book is keyed to a standard mathematics text in the sense that each subsection corresponds to one or a few subsections of the text. For example, the calculus reference is Randolph and Kac's, *Analytic Geometry and Calculus*. There are 91 examples classed under such headings as: inequalities, equation of a line, limits, derivatives, maxima and minima, definite integrals, exponential functions, Taylor's formula, and partial derivatives. As is true throughout, these examples are drawn largely from testing theory, psychophysics, physiological psychology, and learning. Kershner and Wilcox's *The Anatomy of Mathematics* is the text for mathematical foundations. Thirty examples are given, illustrating ideas from the algebra of sets, cartesian products, relations, and functions. The third part on matrix algebra uses Aitken's *Determinants and Matrices* and includes 65 illustrations of such matters as elementary matrix operations, determinants, solutions of linear equations, and linear independence. The final part, devoted to probability theory, refers to Feller's *An Introduction to Probability Theory and its Applications*. Beginning with sample spaces, the 67 examples range over such topics as binomial coefficients, statistical independence, random variables, expectation and variance, and Markov chains.

Considering the rather rapid development of mathematical psychology, one can only hope that the Social Science Research Council will see fit to supplement or revise this useful problem list every five years or so.

R. DUNCAN LUCE

Harvard University

CALVIN S. HALL AND GARDNER LINDZEY, *Theories of Personality*. New York: John Wiley & Sons, Inc., 1957. Pp. xi + 572.

This book is designed to provide a "single source to which the student can turn for a survey of existing theories of personality." It consists of fourteen chapters, twelve of which are devoted to summaries of major (i.e., influential), distinguishable personality theories as identified by Hall and Lindzey and as described by them with the advice and criticism of leading protagonists of the respective theories. The titles of these main chapters

will indicate the range of content of this book: (II) Freud's Psychoanalytic Theory, (III) Jung's Analytic Theory, (IV) Social Psychological Theories: Adler, Fromm, Horney, and Sullivan, (V) Murray's Personology, (VI) Lewin's Field Theory, (VII) Allport's Psychology of the Individual, (VIII) Organismic Theory, (IX) Sheldon's Constitutional Psychology, (X) Factor Theories, (XI) Stimulus-Response Theory, (XII) Rogers' Self Theory, and (XIII) Murphy's Biosocial Theory.

There can be no question but that Hall and Lindzey have performed a valuable service in preparing and making available this material. Heretofore there has been no comparable source, and a noticeable resultant one-sidedness in many individual psychologists' knowledge of personality theory. Furthermore, to the extent that the main chapters have been separately reviewed in advance of publication by those best able to say whether they truly reflect the theories they discuss, we are entitled to regard them as relatively authoritative—at least with respect to matters that are explicitly discussed. At the same time, since all of the chapters are the work of but two authors working in collaboration, there is a unity of presentation and consistency of style that is often lacking in books as eclectic as this one.

A book such as this has many potential uses. Inevitably the requirements suggested by these uses are to some extent in conflict, and it is of interest to see which have won out.

One potential use is as a text. From this point of view one looks for a book that is factually correct, well-organized, clearly written and to some degree inspiring of interest in the subject matter. One also looks for a book that can be flexibly adapted to the idiosyncrasies of a particular teacher's approach to the subject. All these qualities seem to be present.

A second potential use is as a reference book by those already somewhat familiar with the field. On this score one would wish to set higher standards than those achieved for completeness of coverage, both of ideas germane to the various theories and of the surrounding literature. This is not a serious shortcoming in view of the chapters that regularly have appeared in the *Annual Review of Psychology* on personality assessment and related topics.

A third potential use is in the manner of an original contribution. From this point of view it looks as if the authors have consciously refused to make the most of their best opportunity. In the opening and closing chapters, and in the concluding section of each main chapter, Hall and Lindzey come face-to-face with the thorny problems of evaluating personality theory. Admittedly, no stand they might have taken would be likely to please even a majority of their more informed readers. Also, the authors do go through the motions of proposing and discussing over twenty dimensions on which personality theories may be compared, but this leads to the quite predictable observation that any theory looks good by some standards. Such a state of affairs cannot last forever. While it is asserted that "the final crucible for any theory is the world of reality studied under controlled conditions," this is not presented as a principle that is currently capable of untying the Gordian Knot. Neither is it deemed wise "to attempt a synthesis of theories whose empirical utility remains largely undemonstrated."

Under these circumstances perhaps the best strategy for any would-be personality theorist is a perfect illustration of classic scientific method, namely, the isolation of instances in which two personality theories lead to divergent predictions, followed by the unbiased collection and rigorous analysis of pertinent empirical data. A thoughtful reading of Hall and Lindzey's book is bound to suggest many such instances.

D. R. SAUNDERS

Educational Testing Service

G. HERDAN, *Language as Choice and Chance.* Groningen, Netherlands: P. Noordhoff, 1956. Pp. xiii + 356.

Although the publishers announce that this book "aims at providing a systematic

exposition of the quantitative structure of language," the author is more modest and in his preface narrows this down to what he conceives of as the "four main branches of literary statistics: Stylolinguistics, Statistical Linguistics, Information Theory, and Linguistic Duality." The term "literary statistics" is perhaps most apt to describe the contents of the book. The frequent qualification of topics as "linguistic" is in the broad sense of pertaining to language and does not imply that the book makes use of the methods or results of modern descriptive linguistics. With the exception of some material on the distribution of phoneme occurrences, the linguistic data are counts from written sources of the occurrence of words, syllables, letters, and similar elements not regarded as primary in study of structural linguistics. Some of these data are the work of the author and are new; many have been collected from published works.

The statistical problems the author approaches have some intrinsic interest and are doubtless important for the critical study of literary material, but his handling of them is disappointing. For example, to the problem of disputed authorship he brings only abbreviated versions of the techniques used by G. U. Yule in *A Statistical Study of Literary Vocabulary*. These are based on comparing the distribution of nouns by frequency of occurrence in the disputed works with similar distributions from the known works of the contended authors. Yule noted that the means and variances of these (*J*-shaped) distributions increased with sample size (number of running words) and that the variances exceeded the means. By analogy with accident statistics, Yule decided he was dealing with compound Poisson distributions (which are descriptive of the distribution of individuals by number of accidents when risk of accidents differs among individuals). He then showed that the coefficient of variation for the component distributions with respect to their average mean should be independent of sample size and a useful statistic for comparing the noun distributions. When he computed a function of this statistic for the known works of the authors in question, Thomas à Kempis and Gerson, Yule found its value for the disputed work, *De Imitatione Christi*, more in accord with the works of à Kempis. Also, contingency tables between the number of nouns used 1, 2, 3, . . . , n times in the *Imitatione* and 1, 2, 3, . . . , m times in the known works of these authors showed higher association with the works of à Kempis.

Yule's study is highly empirical, as he readily admitted, and not entirely complete or satisfactory from a statistical point of view. It was a work of his later years and explicitly left its more formidable problems for others to solve. G. Herdan presumedly took up the work at this point. He did little more, however, than note that Yule's statistic was only slightly different from the coefficient of variation computed directly from the distribution of nouns. He used this simpler statistic to rework Yule's data and examine some additional material. From a linguistic point of view, Herdan's acceptance of Yule's word counting approach to resolution of disputed authorship seems somewhat hasty. It could be argued, for example, that the range of vocabulary used by a writer depends heavily on the scope of the subject matter discussed. There is danger that the disputed work will be compared with works more similar in content to those of one of the contended authors than those of the other, particularly when the material for comparison is limited. Yule's approach would probably be strongly biased by this error. On the other hand, if the analysis were carried out in terms of classes of constructions at the phrase and sentence level, the results might reflect the formality of the language in addition to the style of the author, but the influence of content should be much reduced. In general, it should be easier to match works by level of formality than by content, suggesting that comparisons in terms of larger constructions would be more reliable than word counts. Even if this is not the case, it seems likely that the influences which determine choice of words are not those which determine choice of larger constructions. If this is true, an analysis at the syntatical level would provide an independent test of conclusions reached on the word level.

The section on "statistical linguistics" in this book is devoted to exhibiting many

examples of frequency distributions of what the author describes as "certain linguistic forms," e.g., phonemes, letters, word length in terms of number of letters and syllables, grammatical forms (classical parts of speech), and metrical units in Latin and Greek hexameter. The author is much impressed by the constancy of these distributions in samples from different tests, and calls it the "basic law of linguistic communication and realization." He makes no attempt to account for the distributions in terms of any model, except to note that the preferred positions of metrical divisions after 3, 5, and 8 syllables in Greek hexameter corresponds to successive terms in that delight of mathematical recreations, the Fibbonaci series (Northrop, E. P. *Riddles in Mathematics: A Book of Paradoxes.* New York: Van Nostrand, 1944).

The introduction to the section on information theory consists of the repetition of Shannon's main results for the discrete case. A species of information measure is then used to characterize a novel sort of relationship between a text in one language and its translation into another. Each word is entered in a two-fold classification according to the number of syllables in the original, and the number of syllables in the corresponding word of the translation. This leads to certain difficulties from free translation, or from German words which correspond to a half a dozen or more in English, or from the habit in Slavic languages not to use articles and to omit the verb "to be." These are resolved by appropriate rules, and the dependencies in the resulting contingency tables are characterized by information measure rather than more conventional measures of association. In these terms the French language turns out to be most similar to English, then German, Czech, and Russian least like English. This, the author says, "reflects the varying degrees of relationship" between the languages. What this relationship might mean in terms of the historical affinities of the languages is not discussed.

The fifth section, "Linguistic Duality," is devoted to various opposing tendencies in language, such as the fact that words of greater frequency tend to be of lesser length, that words pronounced the same tend to acquire several meanings, and concepts or meanings tend to acquire synonymous words, and that freedom and constraint, choice and chance, contribute to determine the sequence of symbols in written language. This section contains data showing that when Chinese ideographs are classified by the number of brush strokes as they are in dictionaries, the number of characters having 1, 2, 3, . . . , 27 brush strokes is distributed much like the number of genera of insects having 1, 2, 3, . . . , n species. The author construes this as evidence that the Chinese dictionary is organized on "taxonomic" principles. If this is the case, then the distribution of words in written text by their frequency of occurrence, of cities by population, and of incomes by size, reflect taxonomic principles also, since the many examples collected by Zipf show that they are also distributed in this way. Herbert Simon has demonstrated that these distributions can be regarded to arise from a stochastic model which yields the limiting form of Yule's well-known distribution of genera by the number of species. This model is of so great generality that we should attach no more profound significance to the fact that certain phenomena conform to it than we do to conformity of other phenomena to the normal distribution.

The final section of this work is a review of large sample statistical methods, through product moment correlation. The treatment is strictly Pearsonian, including the sample size for the denominator of the variance estimator and such topics as the critical ratio test for binomial proportions, and mean square contingency. Since word samples in literary statistics are usually very large, the author feels these methods are adequate.

Much of the interpretation of the statistical results obtained in this book is vague and has tendencies toward the metaphysical in the section on Linguistic Duality. In the opinion of the reviewer, this book cannot be considered a significant contribution to the study of language.

R. Darrell Bock

University of North Carolina

Minutes of the
1958 ANNUAL BUSINESS MEETING
of the
PSYCHOMETRIC SOCIETY

The regular Annual Meeting of the Psychometric Society was held in Washington, D. C. on Tuesday, September 2, 1958. President Frederick Mosteller called the meeting to order at 3:05 P. M.

The minutes of the previous Annual Meeting were read and approved.

On a ballot for the election of two new members of the Council of Directors, Dr. Lloyd G. Humphreys and Dr. Ardie Lubin were elected for a term of three years, ending in 1961.

Dr. John E. Milholland reported for the Membership Committee. The Membership Committee nominated 44 persons as full members and 21 as student members.

It was moved, seconded, and passed that the 21 persons named below be elected as student members.

Vladimir V. Almendinger, Jr., Brighton 35, Massachusetts
Richard F. Arnold, East Lansing, Michigan
Mrs. Joan Hauser Bailey, Van Nuys, California
Mark Philip Bryden, Montreal, Canada
Cherry Ann Clark, South Pasadena, California
Bart B. Cobb, Jr., San Antonio, Texas
Kern William Dickman, Urbana, Illinois
Howard J. Douglas, Lafayette, Indiana
Jean Engler, University of North Carolina, Chapel Hill, North Carolina
Morton P. Friedman, Columbus, Ohio
Arthur H. Hill, University of Minnesota, Minneapolis, Minnesota
George G. Karas, West Lafayette, Indiana
Mrs. Ann S. McColskey, Volusia County Health Unit, Daytona Beach, Florida
Kazuo Nihira, Los Angeles, California
Melvin R. Novick, Chicago, Illinois
LeRoy A. Olson, Madison, Wisconsin
Erich P. Prien, Jr., Cleveland, Ohio
Marvin Snider, Ann Arbor, Michigan
Douglas K. Spiegel, Chapel Hill, North Carolina
Edward E. Ware, Urbana, Illinois
Leonard Wevrick, University of Illinois, Urbana, Illinois

It was moved, seconded, and passed to elect as full members the following 44 individuals.

Joel W. Ager, Jr., Pleasant Ridge, Michigan
Edward F. Alf, Jr., San Diego, California
Eivind Henri Baade, Oslo, Norway
Rudolph G. Berkhouse, Alexandria, Virginia
Allan Birnbaum, Columbia University, New York, New York
Robert F. Boldt, Department of the Army, AGO, Washington, D. C.
Harry Bornstein, Arlington, Virginia
Guido Borasio, Washington University, St. Louis, Missouri
Joan H. Cantor, Peabody College, Nashville, Tennessee
Edward Calvin Carterette, University of California, Los Angeles, California

Robert E. Chandler, Detroit, Michigan
Kenneth E. Clark, University of Minnesota, Mineapolis, Minnesota
William V. Clemans, National Board of Medical Examiners, Philadelphia, Pa.
Dorothy M. Clendenen, The Psychological Corporation, New York, New York
Adriaan D. deGroot, Amsterdam, Holland
Edmund Emil Dudek, U.S. Naval Personnel Research, San Diego, California
Wendell R. Garner, Johns Hopkins University, Baltimore, Maryland
Sten Henrysson, Stockholm, Sweden
Peter A. Holman, Downey, California
Robert Anthony Jones, Redondo Beach, California
Herbert Kaizer, IBM Corporation, Lexington, Massachusetts
D. James Klett, Perry Point, Maryland
Eiichi Komiyama, Kitaku, Tokyo, Japan
Samuel S. Komorita, Vanderbilt University, Nashville, Tennessee
R. Duncan Luce, Cambridge, Massachusetts
Winton Howard Manning, Washington University, St. Louis, Missouri
Philip R. Merrifield, Long Beach, California
Jerome L. Myers, University of Massachusetts, Amherst, Massachusetts
Paul DeLay Nelson, Naval Air Station, Corpus Christi, Texas
Mrs. Nageswari Rajaratnam, Urbana, Illinois
Olav Reiersol, Institutt for Matematiske Tag, Oslo, Norway
James H. Ricks, Jr., The Psychological Corporation, New York, New York
Bryan Boroughs Sargent, III, Knoxville, Tennessee
Paul A. Schwarz, American Institute for Research, Pittsburgh, Pennsylvania
William S. Schwarzbek, General Electric Company, New York, New York
Lee B. Sechrest, Northwestern University, Evanston, Illinois
Robert Seibel, Peekskill, New York
Maynard W. Shelly, Columbus, Ohio
Roger Newland Shepard, Bell Telephone Laboratories, Murray Hill, New Jersey
Walter R. Stellwagon, Syracuse, New York
Patrick Suppes, Stanford University, Stanford, California
James M. Vanderplas, Washington University, St. Louis, Missouri
Charles L. Walter, University of Tennessee, Knoxville, Tennessee
Lawrence K. Waters, Ohio State University, Columbus, Ohio

It was moved, seconded, and passed that the Membership Committee be thanked for their excellent work.

Dr. Irving Lorge reported for the Committee on the Relations between the Psychometric Society and the Psychometric Corporation. A copy of this report is attached. It was moved, seconded, and passed that the report of this committee be accepted with thanks, and that the committee, consisting of Dr. Lorge as Chairman, Dr. Clyde H. Coombs and Dr. John M. Stalnaker, be continued.

President Mosteller asked for a show of hands to indicate the sentiments of the members present on three possibilities of future relationships between the Psychometric Society and the Psychometric Corporation. The first alternative was to continue the present organizational structure. The second alternative was to enlarge the Corporation so that all members of the Society become members of the Corporation. The third proposal was to proceed with the recommendations of the Committee and to take steps to incorporate the Society and to dissolve the Corporation. It appeared to be the sense of the Meeting that the third possibility was the most desirable. It was moved and seconded that the Committee on the Relations between the Psychometric Society and the Psychometric Corporation be instructed to continue to take steps leading to incorporating the Society and the dissolving of the Corporation. The motion was passed unanimously.

It was moved and seconded that the President appoint Dr. William B. Schrader, the Treasurer of the Society, as a member of the Committee on Relations between the Psychometric Society and the Psychometric Corporation. Motion passed.

The motion was made, seconded, and passed that up to $500 be made available for the work of the Committee on Relations between the Psychometric Society and the Psychometric Corporation. Half of this money may be used for expenses in connection with submitting the draft of a new constitution of the Society for the approval of the membership and half for expenses in connection with incorporating the Society. Motion passed.

Dr. J. E. Keith Smith reported for the Program Committee. Of ten abstracts of papers submitted for consideration for presentation at the Annual Meeting, nine were accepted. Three symposia were scheduled, one of which was a proposal submitted by a member and two were developed by the Program Committee. It was moved and seconded that the report be accepted with thanks. Motion passed.

It was moved, seconded, and passed that the Council of Directors obtain information on the affiliation of the Psychometric Society with the American Psychological Association.

The Secretary's report was presented by Dr. Philip H. DuBois. He stated that approximately 20 members, not members of the American Psychological Association, took advantage of the system of registering for the Psychometric Society Meeting by mail. The Secretary's report was accepted with thanks.

It was moved and seconded that the minutes of the Annual Business Meeting be published hereafter in Psychometrika. Motion passed.

It was moved and seconded that a committee be appointed to study special membership categories, including foreign membership and life membership for older individuals. Motion passed.

It was moved and seconded that an auditing committee be appointed to audit the books of the Treasurer. Motion passed.

It was moved and seconded that the President and President-Elect appoint a committee to explore the possibility of special events to celebrate the 25th anniversary of the Society in the year 1960. It is understood that this special committee will not replace the regular Program Committee, but will supplement it. Motion passed.

The report of the Treasurer was presented by Dr. Schrader. A copy is attached. It was accepted with thanks.

Dr. Coombs, reporting for the Elections Committee stated that Dr. Frederic Lord had been elected President of the Psychometric Society for a period of one year beginning October 1, 1958.

The meeting was adjourned at 4:05 P. M.

Philip H. DuBois
Secretary

Report of the
COMMITTEE ON THE RELATIONS BETWEEN
THE PSYCHOMETRIC SOCIETY AND THE PSYCHOMETRIC CORPORATION

September 2, 1958

1. The Committee on the Relations between the Psychometric Society and the Psychometric Corporation reported to the Psychometric Corporation that the most advisable procedure for affecting the merger of the Psychometric Corporation with the Psychometric Society would be to dissolve the Corporation, turn its assets over to the Psychometric Society and to incorporate the Society. The Corporation accepted the Committee report and continued Irving Lorge and John Stalnaker as its representatives on the Joint Committee.

2. During the year, the Committee developed a draft of the Constitution of the Psychometric Society which incorporated the general provisions for the transfer of the assets of the Psychometric Corporation and the continuation of the editorial policies of Psychometrika for an interim period of not more than six years.

3. The Committee has investigated the general procedures for incorporation of the Society as a non-profit organization. It believes, however, that it will need to have legal counsel to determine in what state such incorporation would be most desirable, and, then, to retain counsel to affect the incorporation of the Psychometric Society as a non-profit organization which also will be tax-free.

4. The Committee, therefore, recommends that steps be taken to adopt a new Constitution for the Psychometric Society according to the provisions of Article XII of the current constitution which requires

a) previous approval "by a three-fourths vote of the entire membership of the Council of Directors and the Editorial Council as a whole" and

b) the subsequent approval "by a vote of two-thirds of the Members present at any Annual Meeting or by a two-thirds vote of all Members responding by vote to a mailed ballot."

5. The Committee recommends that the Treasurer of the Society and of the Corporation be added to the joint committee to facilitate the preparation of the Constitution for vote by the Council of Directors and the Editorial Council, to submit the Constitution for a mailed vote of the membership, and to facilitate the designation of the State for incorporation of the Society.

6. The Committee recommended a budget of $250.00 for the preparation and mailing of the proposed new Constitution to the Council of Directors and the Editorial Council and subsequently to the membership, and a budget of $250.00 for legal counsel in the actual incorporation of the Society. The entire sum is to be budgeted by the Psychometric Society.

7. The Committee has given consideration to a number of proposals to develop a suitable memorial to Professor L. L. Thurstone. It will solicit further suggestions for consideration at the next annual meeting of the Society.

Respectfully submitted,

Irving Lorge
Clyde Coombs
John Stalnaker

PSYCHOMETRIC SOCIETY

Statement of Receipts and Disbursements for Fiscal Year Ended June 30, 1958

RECEIPTS (Dues)

Year	Members	Student Members
1958	554	48
1957	41	9
1956	3	
	598	57

	$4,414.00
Received with Dues for Corporation Publications	192.60
Overpayments	.26
Partial Payments	4.60
Total Receipts	$4,611.46

DISBURSEMENTS

Psychometric Corporation (90% of dues)	$3,976.74
Psychometric Corporation (Publications)	192.60
Stationery and Postage	175.70
Secretarial Services	89.27
Bank Charges	8.24
Telephone	9.96
Total Disbursements	$4,452.51

BALANCE

Balance, June 30, 1957	$1,186.25
Receipts, 1957-58	4,611.46
	5,797.71
Disbursements, 1957-58	4,452.51
Balance, June 30, 1958	$1,345.20

PSYCHOMETRIC CORPORATION

Statement of Receipts and Disbursements for Fiscal Year
Ended June 30, 1958

RECEIPTS

Subscriptions (less agency discounts)	$5,656.00
Psychometric Society (90% of dues)	3,976.74
Sale of Back Issues (less discounts)	428.05
Sale of Monographs 5-8 (less discounts)	230.60
Interest on Savings Accounts	271.25
Reprints	662.83
Net overpayments	28.31
	$11,253.78

DISBURSEMENTS

Printing and Mailing Psychometrika Volume 22, No. 2, through 23, No. 1	$6,873.85
Reprints	310.61
Stipend of Managing Editor (7/1/57--6/30/58)	750.00
Stipend of Assistant Managing Editor (7/1/57--6/30/58)	500.00
Stipend of Treasurer (7/1/57--6/30/58)	250.00
Secretarial Services: Editorial Office	800.00
Secretarial Services: Business Office	112.90
Stationery and Postage	185.52
Mailing Back Issues and Monographs	86.28
Refunds	37.80
Miscellaneous	31.85
	$9,938.81

BALANCE AND RESERVES

Balance, June 30, 1957	$5,927.77
Reserve Funds, June 30, 1957	
Englewood Savings and Loan Assn. Englewood, Colorado	3,500.00
Metropolitan Savings and Loan Assn. Los Angeles, California	3,500.00
Total	12,927.77
Receipts, 1957-58	11,253.78
Sum	24,181.55
Disbursements, 1957-58	9,938.81
Remainder	$14,242.74
Balance, June 30, 1958	$ 7,242.74
Reserve Funds, June 30, 1958	
Englewood Savings and Loan Assn. Englewood, Colorado	3,500.00
Metropolitan Savings and Loan Assn. Los Angeles, California	3,500.00
Total, Balance and Reserve Funds	$14,242.74

OBLIGATIONS

Estimated cost of Psychometrika, Vol. 23, Nos. 2-4	
Printing and Mailing	$5,200.00
Stipends (7/1/58--12/31/58)	750.00
Secretarial Services	450.00
Total	$6,400.00
BALANCE AND RESERVES, LESS OBLIGATIONS	$7,842.74

INDEX FOR VOLUME 23

ERRATUM

In Ward, Joe H., Jr., The counseling assignment problem. *Psychometrika*, 1958, **23**, 55–65.

In the center of page 64 the constant following the equals sign should be $\frac{1}{57{,}360}$ rather than $\frac{1}{57{,}630}$.

Psychometrika

A JOURNAL DEVOTED TO THE DEVELOPMENT OF PSYCHOLOGY AS A QUANTITATIVE RATIONAL SCIENCE

THE PSYCHOMETRIC SOCIETY - ORGANIZED IN 1935

VOLUME 23
NUMBER 4
DECEMBER
1958

PSYCHOMETRIKA, the official journal of the Psychometric Society, is devoted to the development of psychology as a quantitative rational science. Issued four time a year, on March 15, June 15, September 15, and December 15.

DECEMBER, 1958, VOLUME 23, NUMBER 4

Published by the Psychometric Society at 1407 Sherwood Avenue, Richmond 5, Virginia. Second-class postage paid at Richmond, Virginia. Editorial Office, Psychometric Laboratory, University of North Carolina, Chapel Hill, North Carolina.

Subscription Price: The regular subscription rate is $14.00 per volume. The subscriber receives each issue as it comes out, and, upon request, a second complete set for binding at the end of the year. All annual subscriptions start with the March issue and cover the calendar year. All back issues but six are available. Back issues are $14.00 per volume (one set only) or $3.50 per issue, with a 20 per cent discount to Psychometric Society members. Members of the Psychometric Society pay annual dues of $7.00, of which $6.30 is in payment of a subscription to *Psychometrika*. Student members of the Psychometric Society pay annual dues of $4.00, of which $3.60 is in payment for the journal.

Application for membership and student membership in the Psychometric Society, together with a check for dues for the calendar year in which application is made, should be sent to

LLOYD G. HUMPHREYS
Department of Psychology
University of Illinois
Urbana, Illinois

Payments: All bills and orders are payable in advance.

Checks covering membership dues should be made payable to the *Psychometric Society*.

Checks covering regular subscriptions to *Psychometrika* (for nonmembers of the Psychometric Society) and back issue orders should be made payable to the *Psychometric Corporation*. All checks, notices of change of address, and business communications should be sent to

WILLIAM B. SCHRADER, Treasurer, Psychometric Society and Psychometric Corporation
Educational Testing Service
P.O. Box 592
Princeton, New Jersey

Articles on the following subjects are published in *Psychometrika:*

(1) the development of quantitative rationale for the solution of psychological problems;
(2) general theoretical articles on quantitative methodology in the social and biological sciences;
(3) new mathematical and statistical techniques for the evaluation of psychological data;
(4) aids in the application of statistical techniques, such as nomographs, tables, worksheet layouts, forms, and apparatus;
(5) critiques or reviews of significant studies involving the use of quantitative techniques.

The emphasis is to be placed on articles of type (1), insofar as articles of this type are available.

(Continued on the back inside cover page)

In the selection of the articles to be printed in *Psychometrika*, an effort is made to obtain objectivity of choice. All manuscripts are received by one person, who first removes from each article the name of contributor and institution. The article is then sent to three or more persons who make independent judgments upon the suitability of the article submitted. This procedure seems to offer a possible mechanism for making judicious and fair selections.

Prospective authors are referred to the "Rules for Preparation of Manuscripts for *Psychometrika*," contained in the March, 1958 issue. Reprints of these "Rules" are available from the managing editor upon request. A manuscript which fails to comply with these requirements will be returned to the author for revision.

Authors will receive 100 reprints without covers, free of charge.

Manuscripts for publication in *Psychometrika* should be sent to

LYLE V. JONES, Managing Editor, *Psychometrika*
Psychometric Laboratory, Univ. of North Carolina
Chapel Hill, North Carolina

Material for review in *Psychometrika* should be sent to

JOHN B. CARROLL, Review Editor, *Psychometrika*
7 Kirkland Street
Cambridge 38, Massachusetts

The officers of the Psychometric Society for the year October 1958 through September 1959 are as follows: *President:* Frederic M. Lord, Educational Testing Service, P. O. Box 592, Princeton, New Jersey; *Secretary:* Philip H. Dubois, Dept. of Psychology, Washington University, St. Louis 5, Missouri; *Treasurer:* William B. Schrader, Educational Testing Service, P. O. Box 592, Princeton, New Jersey.

The Council members, together with dates at which terms expire, are as follows: Lloyd G. Humphreys, 1961; Ardie Lubin, 1961; Harold P. Bechtoldt, 1960; J. B. Carroll, 1960; T. W. Anderson, 1959; William G. Mollenkopf, 1959.

The *Psychometric Monographs Committee* is composed of Frederick B. Davis, *Chairman;* Harold Gulliksen, Paul Horst, and Frederic Kuder. Manuscripts and correspondence for this series should be addressed to

FREDERICK B. DAVIS, Chairman, Psychometric Monographs Committee,
Hunter College, 695 Park Avenue, New York 21, N. Y.

PRINTED BY THE WILLIAM BYRD PRESS, INCORPORATED, RICHMOND, VA.